MODELLING AND FORECASTING LUNG CANCER INCIDENCE AND MORTALITY IN SAUDI ARABIA

Salford Business School

University of Salford, Manchester, UK

PhD Thesis

By

Salem Mubarak Alzahrani

Submitted in Partial Fulfilment of the Requirements of the Degree of Doctor of Philosophy, November 2016

Declaration

I declare that this thesis is my original work. No portion of this work has been previously submitted for another degree or qualification of this or any other University.

Acknowledgments

I am heartily thankful to my supervisor, Prof. Phil Scarf, whose supervision, encouragement and support from the preliminary to the concluding level enabled me to develop an understanding of the subject. I am sure it would have not been possible without his help. I would also thank the Saudi Cancer Registry, the Ministry of Health, the Central Department of Statistics & Information, the Department of Tobacco Control Program in the Ministry of Health, National Cancer Information Service and King Faisal Specialist Hospital and Research Centre in Riyadh for providing me with published data. It is a pleasure to thank those who made this thesis possible especially my parents, waif, daughter, family and friends in the UK who gave me the moral support I required. Lastly, I offer my regards and blessings to all of the University of Al-Baha, the Ministry of Education and the Royal Embassy of Saudi Cultural Bureau in the UK who supported me in many respects during my programme and the completion of the project.

Abstract

The aim of this research is to forecast the rates of lung cancer incidence and mortality in the Kingdom of Saudi Arabia using data on lung cancer diagnosis between 1994 and 2009. Lung cancer data, including incidence and mortality, were obtained from Saudi Cancer Registry at the Ministry of Health. The Central Department of Statistics & Information at the Ministry of Planning also provided data on person characteristics, such as age, gender and ethnicity. These data serve as a basis for modelling the effect of gender, ethnicity, and age at diagnosis, and region on incidence and mortality. For comparison of incidence and mortality rates between region and over time, standardised rates are used in this thesis, based on a hypothetical standard population, in our case the world standard population. We use several modelling approaches. The first part of the analysis uses two approaches. The first approach concentrates on Box–Jenkins methodology, and the second approach uses dynamic regression modelling including both finite and infinite lag models to forecast lung cancer incident cases. The second part focuses on age-period-cohort modelling including both incidence and mortality rates of lung cancer, and using two methodological approaches, namely spline functions and Bayesian dynamic models, for the incidence and mortality respectively. Lung cancer is rarely diagnosed in people under 30 years of age in Saudi Arabia, but incidence rises sharply thereafter peaking in the 65-69 years age group. Males have a 79% greater incidence rate of lung cancer than females across the entire dataset when adjusting for the other effects. The average age standardised incidence rate in 2009 was 3.8 per 100,000 population whereas the average age standardised mortality rate was 1.9 per 100,000 population in the same year. The highest number of cases of lung cancer were reported in the Western region at 187 and in Riyadh at 144 cases and the majority of cases were diagnosed in winter (December - March). The forecast incidence rate of lung cancer is expected to decrease in men but to increase in women over the next ten years. This is perhaps due to the increase in the proportion of female smokers. The male age standardised rate of lung cancer incidence is forecast to fall from 4.6 in 2010 to 2.4 per 100,000 by 2020, whereas the female age standardised rate is forecast to increase from 2.0 in 2010 to 2.2 per 100,000 by 2020. On the other hand, the overall mortality rate of lung cancer (with 95% credible interval shown) is forecast to increase to 2020 from 1.8 (1.61, 1.94) in 2010 to 3.04 (0.13, 5.94) per 100,000 population. Age has a strong association with lung cancer mortality, suggesting age-related causes such as accumulative exposures to smoking over time may be the main reason for increasing lung cancer mortality in Saudi Arabia. This is the first study to forecast lung cancer incidence and mortality in Saudi Arabia. It will help the Saudi Arabian Ministry of Health to understand the rate of future lung cancer incidence and mortality and the overall effects of the population classes, and to plan healthcare provision accordingly. The data are limited because the Saudi Cancer Registry has only been in existence since 1992. Therefore, we can expect the precision of forecasts to improve as further data are collected.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

1.1. Background Information

Cancer is a major health challenge. Globally, the estimated number diagnosed with cancer is approximately 14.1 million people per year and mortality is 8.2 million deaths per year (Ferlay et al, Global Cancer, 2012; IARC, 2013). These figures are set to rise to 26.4 million and 13.2 million by 2030 (Boyle P, et al. World Cancer Report, 2008).

At the beginning of the 20th century lung cancer was a very rare disease. The increase was first recognized in autopsy research (De Vries VM, 1927). Since World War II, rates in the Western world have increased dramatically and lung cancer could be called 'one of the epidemics' of the 20th century. Nowadays lung cancer is the first or second most frequent tumor type among men and third or fourth among women (World Health Organization, Media Center, 2015).

The Kingdom of Saudi Arabia (KSA) is the largest country of the Arabian Peninsula and the second-largest country in the Arab world. It extends from the Red Sea in the west to the Arabian Gulf in the east with approximately 2,150,000 square kilometers in land area. KSA is divided into 13 administrative regions. In 2010, the population was approximately 27 million (Figure 1.1).

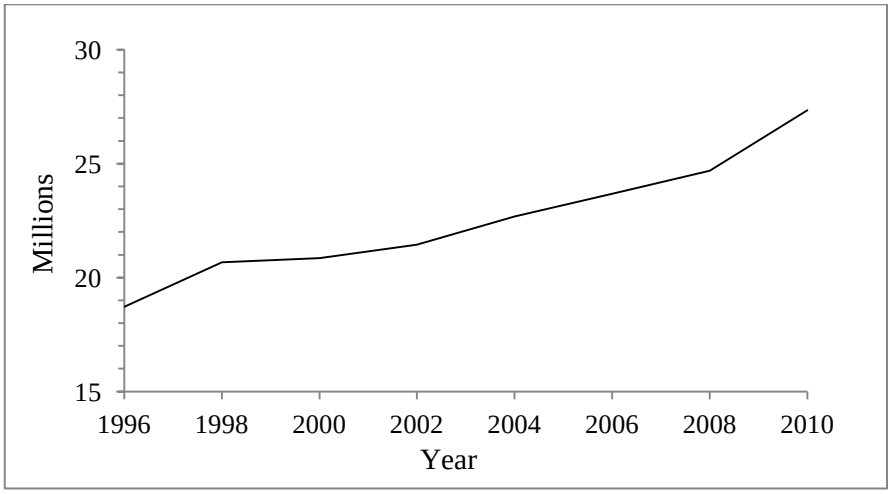


Figure 1.1: Population of Saudi Arabia from 1996 to 2010.

Tobacco smoking is the most important risk factor for cancer, causing 20% of the world mortalities and more than 70% of global lung cancer mortalities (WHO, Media

Center, 2014; Cancer research UK, 2014). Globally, three people die every minute from lung cancer according to WHO (Elsayed et al., 2011). Lung cancer is a multifactorial disease – that is, many factors work together to either cause or prevent lung cancer. Other risk factors include genetic risk, age, effects of past cancer treatment, exposure to asbestos, radon gas and – in very rare cases – substances such as uranium, chromium, nickel, and polycyclic hydrocarbons (Alberg and Samet, 2003). Lung cancer is not infectious.

In the Kingdom of Saudi Arabia (KSA), the amount of imported tobacco has increased dramatically in recent years (Figure 1.2). This suggests there will be a serious problem with lung cancer in the future. The sharp decrease in the period 1991 and 1996 coincides with the Gulf War.

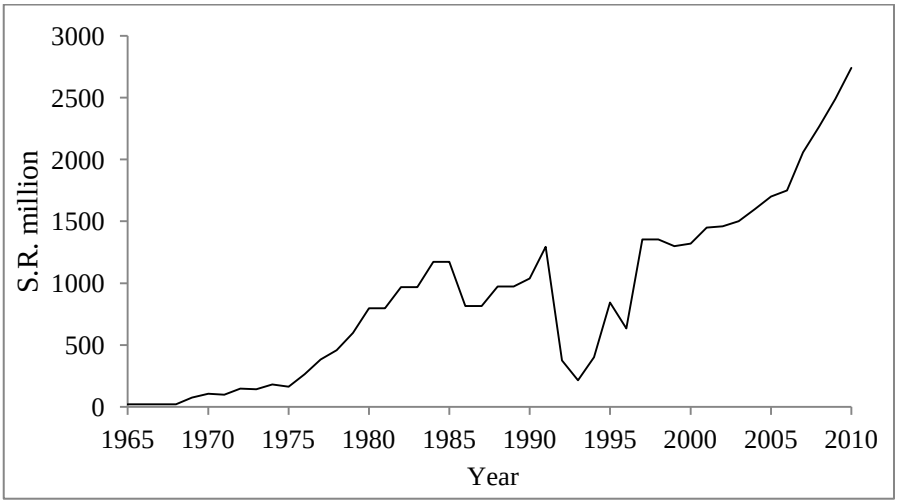


Figure 1.2: Value of tobacco imported by Saudi Arabia (1 S.R. = $3.75).

Lung cancer is estimated as the seventh most common cancer in Gulf Countries (Gulf Cooperation Council, 2011). In 2007, the estimated lung cancer cases were 4600 and accounted for 5% of all cancers. In Gulf Countries, the average age standardised rates reported were 7.0 per 100,000 population for males and 2.1 per 100,000 for females. Lung cancer seems to be higher among men than women in the Gulf Countries. The highest ASR was in Bahrain at 31.1 and 10.7 per 100,000 population in male and female, respectively. This was followed by Kuwait and Qatar. The lowest ASR was reported in Saudi Arabia at 5.6 and 1.6 per 100,000 population for male and female, respectively (Gulf Cooperation Council, 2011). Lung cancer ranked in the seventh position with around 490 cases at 3.9% of all cancers (Al-Eid, Saudi Cancer Incidence Report, 2009). This percentage could increase in future according to the chairman of the Saudi cancer registry.

1.2. Aims and Objectives

The aim of this thesis is to use statistical methods to model temporal trends of lung cancer in the Kingdom of Saudi Arabia (KSA) and to predict cancer incidence and mortality up to 2020. We develop trend models for the period 1994-2009 for different age groups for short and medium term predictions. In so doing we aim to produce forecasts of number of cases by that use additional information available about male and female smoking prevalence and other covariates. In addition, we aim to describe the broad picture of the future lung cancer burden in KSA against which progress in implementing the National Health Service (NHS) Cancer Plan will be measured.

Projecting the burden of cancer is important for evaluating prevention strategies and for administrative planning at cancer facilities. Health and planning officials need to plan treatment and care. In fact, assuming that the current rates will remain the same is often inaccurate.

1.3. Methodology

We study lung cancer incidence in Saudi Arabia between 1994 and 2009. Lung cancer incidence and mortality data were obtained from Saudi Cancer Registry (SCR). The Central Department of Statistics & Information (CDS) provided data on person characteristics, such as age, gender, and ethnicity from 1994 to 2009.

In the first part of this research, the incidence of lung cancer are modelled and predicted using Box-Jenkins methodology and dynamic regression models. Box-Jenkins methodology fits non-seasonal Autoregressive Integrated Moving Average (ARIMA) models and seasonal ARIMA (SARIMA) models. Dynamic regression models would involve more general autoregressive AR(∞) processes such as AR(1), distributed lag models (DLMs), and polynomial distributed lag models (PDLs). We try to find new approaches to evaluate the robustness of the results, using autoregressive polynomial distributed lag models (ARPDLs). Thus, the ultimate purpose of dynamic modelling is to estimate consistent individual (short run) and cumulative (long run) trends of lung cancer cases over the period 1994-2009 per month and to forecast over the period 2010-2012. The second part of this research concentrates first on the age-period-cohort (APC) modelling using the spline functions for the incidence rates and second on Bayesian dynamic APC modelling for the mortality rates of lung cancer. We forecast the rates of lung cancer incidence and mortality up to 2020 for the population of Saudi Arabia using population projections from the United Nations (2012) for the years 2010 to 2020 using lung cancer

incidence and mortality data from the Saudi Cancer Registry (SCR) for the years 1994 to 2009.

We model the incidence rate of lung cancer using a version of the age-period-cohort model with recommended modifications that was developed and tested in Stata Journal articles (Rutherford et al., 2010 and Sasieni, 2012). In the age-period-cohort setting, we fit spline functions to each of the three components of age, period, and cohort. Constraints need to be made because of the lack of identifiability of the model. The identifiability issue stems from the fact that there is an exact relationship between the variables.

Rutherford et al. (2010) described an APC command called `apcfit` and illustrated how to fit age-period-cohort models when not making predictions. Potentially, an update by Sasieni (2012) made predictions easier from `apcfit` command. The extension to making the predictions involves a little care in setting up the data and making the knot selection with simple assumptions of linearity beyond the boundary knots. Using the restriction of the cubic splines being linear beyond the boundary knots, we were able to make better predictions in the magnitude of the rates, the variation by age, and time trends in the rates. We arrange the data in one-year intervals from 1994 to 2009 and 5-year age groups from 0-4 years to 75+ years. We obtain parameters by means of a maximum likelihood procedure. We add covariates to the models in order to obtain the best-fitting model using the model selection criteria. In this analysis, various combinations of covariates such as gender, race, smoking prevalence by gender, price of imported tobacco, consumption of tobacco per 1000 tons were used. In addition, five created regions (north, south, east, west, central) from the whole 13 administrative regions of Saudi Arabia were added to assess the performance of the final model.

In the Bayesian dynamic APC modelling, we follow the strategy proposed by Held and Rainer (2001) and Shuichi et al. (2008) by using a dynamic age-period-cohort model to smooth age, period and cohort trends and to extrapolate N future periods and cohorts. We model lung cancer mortality trends through specific smoothing of model parameters since our lung cancer mortality data are sparse (many of zero counts). According to Knorr-Held and Rainer (2001), a second order random walk (RW2) has been assumed for age, period and cohort effects to reduce the variation of parameter estimates. We calculate lung cancer mortality rates using the population of Saudi Arabia from 1994-2009. We standardize both the incidence and mortality rates using the world standard population. We arrange the data in one-year intervals from 1994 to 2009 and 5-year age groups from 25-29 years to 75+ years. Since there are fewer observations for the earliest and most recent

cohorts this may lead to less precision in the estimates of these cohorts. The form of models fall into the class of generalized linear models with the number of lung cancer incidence and mortality follow a Poisson distribution. The posterior distributions of the hyper-parameters are obtained by using Markov Chain Monte Carlo (MCMC) techniques. To achieve better smoothing of the parametric effects, we introduce an adaptive precision parameter (K_s) for each prior distribution of age, period and cohort as suggested by Cleries et al. (2010). In the Bayesian analysis, convergence diagnostics and model selection criteria are used to compare between nested models and select the best-fitting model.

The data in this thesis are analyzed using statistical software packages Minitab, Stata13, Eviews8, R, and R2WinBUGS.

1.4. Justification

Recently, cancer has become the top priority of the government of Saudi Arabia because of its increase in the country. Therefore, effort must be made to reduce and prevent the increase of cancer incidence and mortality in Saudi Arabia.

In 2009, the percentage of males and females smokers in Saudi Arabia has been estimated to be around 20.8% for males and 5.8% for females for the population aged 16 and over (Ministry of Health, 2009). This implies around 3,775,400 million adult cigarette smokers in KSA. Thus, efforts need to be made to reduce the prevalence of smoking since tobacco is responsible for around 70% of lung cancer mortality (World Health Organization, Media Center, 2015). In addition to the human toll of cancer, the financial cost of cancer is substantial. The direct costs include payments and resources used for treatment and the indirect costs include the loss of economic output due to days missed from work.

1.5. Structure of the Thesis

This thesis is composed of eight chapters. Chapter one is this introduction, where an overview of the project is given and includes the aims and objectives, methodology, and justification. Chapter two is composed of a brief literature review on disease incidence, forecasting lung cancer incidence in developed and developing countries among other things, a review of time series methods, a review of dynamic regression models, a review of age–period–cohort (APC) models, methods for quantification of incidence and mortality, and forecasting methods. Chapter three presents all the data summary. The fourth chapter presents analysis of Box–Jenkins methodology for modelling and forecasting lung cancer incidence. The fifth chapter presents analysis of first order

autocorrelated, distributed lag models and their one-step ahead forecasts, polynomial distributed lag models (PDLs), and autoregressive polynomial distributed lag models (ARPDLs). Chapter six is composed of APC modelling and predictions to 2020 for the incidence using spline functions. Chapter seven presents Bayesian dynamic APC modelling and predictions to 2020 for mortality. Chapter eight presents conclusions, recommendations and future research. Apart from the first and the last chapters, each chapter is provided with a brief summary.

CHAPTER 2

LITERATURE REVIEW

2.1. Cancer Incidence and Mortality

WHO reports that future death rates can be reduced with timely diagnosis, regular screenings, and early treatment of cancers. In 2012, the incidence and mortality cases in the economically developed countries were about 6.1 and 3.0 million, respectively. Whereas, the incidence and mortality cases in economically developing countries were about 8.1 and 5.3 million, respectively (Ferly et al, Global Cancer, 2012). This increase of incidence and mortality cases is simply because of the growth and ageing of the population (American Cancer Society, 2011). Cancer involves more than 100 types of cancers with different etiologic factors and treatment. In 2012, the majority of cancer cases were diagnosed in Eastern Asia at 4,145,000 cases in both males and females. This was followed by Northern America, South-Central Asia and Western Europe at around 1,786,400, 1,514,000 and 1,110,300 cancer cases, respectively (Ferly et al, Global Cancer, 2012). It has been estimated that the number of deaths in Europe is projected to increase by 11% in 2015, compared to the 2000 level (Quinn et al., 2003). It is estimated that in Europe alone, one in three people will be affected by cancer in their lifetime (World Cancer Report, 2003). In Western Asia, the estimated number of cancer incidence was 317,600 cases and the mortality was almost 189,400 cases (Ferly et al, Global Cancer, 2012). In Saudi Arabia, in 2007, the estimated cancer cases were 70,000 with 35,100 cases among males and 34,900 cases among females, compared to 2004 when there were 45,500 cancer cases for both genders (Al-Amadi, K. and Al-Ameri, A., 2011).

2.2. Lung Cancer Incidence and Mortality

In 2012, globally the estimated lung cancer impact was approximately 1.89 million cases and 1.59 million deaths. The numbers of incidence and mortality cases in the developed countries were about 758,000 and 627,000, respectively. In addition, the numbers of cases of incidence and mortality in the developing countries were about 1.1 million and 963,000 respectively (Ferly et al, Global Cancer, 2012). Tobacco smoking is the most important risk factor for lung cancer causing 20% of the world mortalities and more than 70% of global lung cancer mortalities (WHO, Media Center, 2014; Cancer Research UK, 2014).

Globally, the highest age standardised rate (ASR) of lung cancer was among males in Central and Eastern Europe at 53.5 per 100,000 population. This was followed by Eastern

Asia males at 50.4 per 100,000 population. Also, it was high among males in Southern Europe, Western Europe, and North America at approximately 46, 45, and 44 per 100,000 population respectively. Notably, the lowest ASR was in Middle and Western Africa at 2.0 and 1.7 per 100,000 respectively. In females, age standardised rates (ASR) were high in North America and Northern Europe at almost 33.8 and 23.7 per 100,000 population respectively. The lowest ASR was reported again in Middle and Western Africa at 1.1 and 0.8 per 100,000 respectively (Ferly et al, Global Cancer, 2012).

2.3. Time Series Forecasting Models

2.3.1. Introduction

In this section, we highlight brief literature review on time series methods including key publications in other journals. We provide a selective guide to the literature on time series forecasting, covering more than four decades. The proportion of papers that concern time series forecasting has been fairly stable over time. We also review key papers and books published elsewhere that have been highly influential to various developments in the field, but of course the list is far from exhaustive.

The main aim of time series modelling is to carefully collect and rigorously study the past observations of a time series to develop an appropriate model which describes the inherent structure of the series. This model is then used to generate future values for the series, i.e. to make forecasts. Time series forecasting thus can be termed as the act of predicting the future by understanding the past (Raicharoen et al., 2003). Due to the indispensable importance of time series forecasting in numerous practical fields such as business, economics, finance, science and engineering, etc. (Tong, 2003 and Zhang, 2003; 2007), proper care should be taken to fit an adequate model to the underlying time series. A lot of efforts have been done by researchers over many years for the development of efficient models to improve the forecasting accuracy. As a result, various important time series forecasting models have been evolved in literature.

2.3.2. Definition of A Time Series

A data set containing observations on a single phenomenon (or variable) observed over multiple time periods is called time series. It is mathematically defined as a set of vectors $x(t)$, $t = 0, 1, 2, ...$ where t represents the time elapsed (Cochrane, 1997; Hipel and McLeod, 1994; Raicharoen et al., 2003). The variable $x(t)$ is treated as a random variable. The measurements taken during an event in a time series are arranged in a proper chronological

order. In time series data, both the *values* and the *ordering* of the data points have meaning. Although the ordering is usually through time, particularly in terms of some equally spaced intervals, the ordering may also be taken through other dimensions such as space (Wei, 1990).

A time series containing observations of a single variable is termed as univariate, whereas if observations of more than one variable are considered, it is termed as multivariate. A time series can be continuous or discrete. Continuous time series are generally recorded steadily and instantaneously whereas discrete time series contain observations measured at sequential integer values of the variable time. For example temperature readings, flow of a river, concentration of a chemical process, an oscillograph records of harmonic oscillations of an audio amplifier etc. can be recorded as a continuous time series. On the other hand population of a particular city, production of a company, exchange rates between two different currencies, and rainfall accumulations measured at a regular interval may represent discrete time series. Usually in a discrete time series the consecutive observations are recorded at equally spaced time intervals such as hourly, daily, weekly, monthly or yearly time separations. As mentioned in (Hipel and McLeod, 1994), the variable being observed in a discrete time series is assumed to be measured as a continuous variable using the real number scale. Furthermore a continuous time series can be easily transformed to a discrete one by merging data together over a specified time interval. This thesis examines raw data and summary statistics measured at regular intervals over time, for which time series analysis is most appropriate.

Analysis of time series has been a part of statistics for long. Some methods have also been developed for its analysis to suit the distinct features of time series data, which differ both from cross section and panel or pooled data. Various approaches are available for time series modelling. Some of the tools and models which can be used for time series analysis, modelling and forecasting are briefly discussed.

2.3.3. Time Series Models and Components

A time series is a set of values of a particular variable that occur over a period of time in a certain pattern. The time series movements of such chronological data can be decomposed into the most common patterns as increasing or decreasing *trend*, *cyclical*, *seasonal (periodic)*, and *irregular fluctuations* (Bowerman et al., 2005). In some series, one or two of these components may overshadow the others. A basic assumption in any time series analysis and modelling is that some aspects of the past pattern will continue to remain in

the future. For detailed discussion of the four main time series components, see Bowerman et al., (2005).

A time series is non-deterministic in nature, i.e. we cannot predict with certainty what will occur in future. Generally a time series $\{x(t), t = 0,1, 2,...\}$ is assumed to follow certain probability model (Cochrane, 1997) which describes the joint distribution of the random variable x_t. According to Hipel and McLeod (1994), the mathematical expression describing the probability structure of a time series is a stochastic process. Thus the sequence of observations of the series is actually a sample realization of the stochastic process that produced it.

A usual assumption is that the time series variables x_t are independent and identically distributed following the normal distribution. However as mentioned in Cochrane, (1997), an interesting point is that time series are in fact not exactly independent and identically distributed; they follow more or less some regular pattern in long term. For example if the temperature today of a particular city is extremely high, then it can be reasonably presumed that tomorrow's temperature will also likely to be high. Hence, if time series models are put to use, say, for instance, for forecasting purposes, then they are especially applicable only in the short term.

Exponential smoothing methods originated in the 1950s and 1960s with the work of Brown (1959, 1963), Holt (1957, reprinted 2004) and Winters (1960). Pegels (1969) provided a simple but useful classification of the trend and the seasonal patterns depending on whether they are additive (linear) or multiplicative (nonlinear). Muth (1960) was the first to suggest a statistical foundation for simple exponential smoothing (SES) by demonstrating that it provided the optimal forecasts for a random walk plus noise. Further steps towards putting exponential smoothing within a statistical framework were provided by Box & Jenkins (1970, 1976), Roberts (1982) and Abraham and Ledolter (1983,1984), who showed that some linear exponential smoothing forecasts arise as special cases of ARIMA models. However, these results did not extend to any nonlinear exponential smoothing methods. Forty years ago, exponential smoothing methods were often considered a collection of ad hoc techniques for extrapolating various types of univariate time series. Although exponential smoothing methods were widely used in business and industry, they had received little attention from statisticians and did not have a well-developed statistical foundation. A decent account on exponential smoothing methods has been given in Makridakis et al.,(1998).

2.3.4. Models of Stationary Processes

The concept of stationarity of a stochastic process can be visualized as a form of statistical equilibrium (Hipel and McLeod, 1994). The statistical properties such as mean and variance of a stationary process do not depend upon time. It is a necessary condition for building a time series model that is useful for future forecasting. A time series is said to be stationary if its underlying generating process is based on a constant mean and constant variance with its autocorrelation function (ACF) essentially constant through time. This means that different subsets of a time series sample will typically have time independent means, variances and autocorrelation functions that do not differ significantly.

A statistical test for stationarity or test for unit root has been proposed by Dickey and Fuller (1979). The test is applied for the parameter ρ in the auxiliary regression

$$\Delta_1 y_t = \rho y_{t-1} + \alpha_1 \Delta_1 y_{t-1}$$

where Δ_1 denotes the difference operator i.e. $\Delta_1 y_t = y_t - y_{t-1}$.

The relevant null hypothesis is $\rho = 0$ i.e. the original series is non stationary and the alternative is $\rho < 0$ i.e. the original series is stationary. Differencing is usually applied until the acf shows an interpretable pattern with only a few significant autocorrelations.

As mentioned in (Box & Jenkins, 1970 ; Hipel and McLeod, 1994), stationarity is a mathematical idea constructed to simplify the theoretical and practical development of stochastic processes. To build a suitable time series model for future forecasting, the underlying time series is expected to be stationary. Unfortunately this is not always the case. As stated by Hipel and McLeod (1994), the greater the time span of historical observations, the greater is the chance that the time series will exhibit non stationary characteristics. However for relatively short time span, one can reasonably model the series using a stationary stochastic process. Usually time series with trend or seasonal patterns are non stationary in nature. In such cases, differencing and power transformations are often used to remove the trend and to make the series stationary.

2.3.5. Box-Jenkins Methodology

Generally, early attempts to study time series particularly in the nineteenth century were characterized by the idea of a deterministic world. It was the major contribution of Yule (1927) who launched the notion of stochasticity in time series by postulating that every time series can be regarded as the realization of a stochastic process. Based on this simple idea, a number of time series methods have been developed since then.

Autoregressive (AR) models were first introduced by Yule in 1926. They were consequently supplemented by pioneers such as Slutsky who in 1937 formulated moving average (MA) schemes. Wold (1938), first combined both AR and MA schemes and showed that ARMA processes can be used to model all stationary time series as long as the appropriate order of *p*, the number of AR terms, and *q*, the number of MA terms, were appropriately specified. This means that any series can be modelled as a combination of past values and/or past errors. Wold's decomposition theorem led to the formulation and solution of the linear forecasting problem by Kolmogorov (1941). Since then, a considerable body of literature in the area of time series dealing with the parameter estimation, identification, model checking, and forecasting has appeared (see, for example, Newbold, 1983) for an early survey.

Box and Jenkins (1970, 1976) first integrated the existing knowledge, formulated the concepts of ARIMA and popularised the use of ARMA models. Moreover, they developed a coherent, versatile approach for model-building through the following:

i. providing guidelines for making the series stationary in both its mean and variance
ii. suggesting the use of autocorrelations and partial autocorrelation coefficients for determining appropriate values of *p* and q (and their seasonal equivalent *P* and *Q* when the series exhibited seasonality)
iii. providing a set of computer programs to help users identify appropriate values for *p* and *q*, as well as *P* and *Q*, and estimate the parameters involved
iv. once the parameters of the model were estimated, a diagnostic check was proposed to determine whether or not the residuals were white noise, in which case the order of the model was considered final (otherwise another model was entertained in (ii) and steps (iii) and (iv) were repeated). If the diagnostic check showed random residuals then the model developed was used for forecasting or control purposes assuming of course constancy, that is that the order of the model and its non-stationary behaviour, if any, would remain the same during the forecasting, or control, phase.

The approach proposed by Box and Jenkins came to be known as the Box-Jenkins methodology to ARIMA models, where the letter "I", between AR and MA, stood for the word "Integrated". For seasonal time series, a variation of ARIMA, namely, the *Seasonal Autoregressive Integrated Moving Average* (SARIMA) *(*Box and Jenkins, 1970; Hipel and McLeod, 1994; Hamzacebi, 2008) model is used. The *Autoregressive Fractionally Integrated Moving Average (ARFIMA)* (Galbraith and Zinde-Walsh, 2001) model

generalizes ARMA and ARIMA models. ARIMA model and its different variations are based on the famous Box-Jenkins principle (Box and Jenkins, 1970; Zhang, 2003) and these are broadly known as the Box-Jenkins models. In the 1970s, Box-Jenkins methodology became highly popular among academics especially when it was proved through empirical studies using real data that they could outperform the large and complex econometric models, popular at that time, in a variety of situations (Cooper, 1972; Nelson, 1972; Elliot, 1973; Narasimham et al., 1974; McWhorter, 1975; for a survey see Armstrong, 1978). An excellent discussion of various aspects of this approach is given in Box et al. (2007).

In this section, we will present the steps in the ARIMA (and/or SARIMA) methodology. The methodology put forth by Box and Jenkins will be demonstrated with real lung cancer data in another chapter, since it uses several time series procedures.

2.3.6. The Univariate ARIMA Model

The success of the Box-Jenkins methodology is founded on the fact that the various models can, between them, mimic the behaviour of diverse types of series and do so adequately without usually requiring very many parameters to be estimated in the final choice of the model. Univariate models are sometimes referred to as non-causal models. Although our focus is on forecasting, univariate Box-Jenkins models (often referred to as ARIMA models) are often useful for simply explaining the past behaviour of a single data series, for whatever reason one may want to do so. In general, a univariate time series will reflect the reality in which observations occurring close in time have a greater relationship than observations that are farther apart, looking only at the single variable. It is the purpose, therefore, of univariate time-series methods to statistically measure the degree of this relationship.

Notwithstanding, model selection in the mid-sixties was very much a matter of researcher's judgment as there was no algorithm to specify a model uniquely. Since then, many techniques and methods have been suggested including Akaike's information criterion (AIC), Akaike's final prediction error (FPE), and the Bayes information criterion (BIC). Most often, these criteria minimise (in-sample) one-step-ahead forecast errors with a penalty term for overfitting. FPE has also been generalized for multi-step-ahead forecasting (for more details, see Bhansali, 1996, 1999), but this generalization has not been utilized by applied workers. This also seems to be the case with criteria based on cross-validation and split-sample validation (see for example, West, 1996) principles,

making use of genuine out-of-sample forecast errors (Pena & Sanchez, 2005) for a related approach worth considering.

2.3.7. Non-Seasonal ARIMA Models

A stochastic model for non-seasonal series are called *Autoregressive Integrated Moving Average model*, denoted by ARIMA (*p, d, q*). Here *p* indicates the order of the autoregressive part, *d* indicates the amount of differencing, and *q* indicates the order of the moving average part. If the original series is stationary, *d* = 0 and the ARIMA models reduce to the ARMA models.

2.3.8. The Autoregressive Moving Average (ARMA) Models

An ARMA(*p, q*) model is a combination of AR(*p*) and MA(*q*) models and is suitable for univariate time series modeling. In an AR(*p*) model the future value of a variable is assumed to be a linear combination of *p* past observations and a random error together with a constant term. Mathematically the AR(*p*) model can be expressed as (Hipel and McLeod, 1994; Lee, Econs 413, Lecture 4):

$$x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \omega_t = \phi_0 + \sum_{i=1}^{p} \phi_i \, x_{t-i} + \omega_t$$

where x_t and ω_t are respectively the actual value and random error (or random shock) at time period *t* , $\phi_i (i = 1, 2, \dots, p)$ are model parameters to be estimated, ϕ_0 is a constant and *p* is the order of the model. Sometimes the constant term is omitted for simplicity. Usually Yule-Walker equations (Hipel and McLeod, 1994) are used for estimating parameters of an AR process using the given time series.

Whereas an AR(*p*) model regress against past values of the series, an MA(*q*) model uses past errors as the explanatory variables. The MA(*q*) model is given by (Cochrane, 1997, 2005; Hipel and McLeod, 2005):

$$x_t = \mu + \theta_1 \omega_{t-1} + \theta_2 \omega_{t-2} + \cdots + \theta_q \omega_{t-q} + \omega_t = \phi_0 + \sum_{j=1}^{q} \theta_j \omega_{t-j} + \omega_t$$

where μ is the constant mean of the process, $\theta_j (j = 1, 2, \dots, q)$ are the model parameters to be estimated and *q* is the order of the model. The error terms are assumed to be a white noise process, i.e. a sequence of independent and identically distributed random variables with zero mean and constant variance σ^2 (Cochrane, 1997, 2005; Hipel and McLeod, 2005). Generally, the random shocks are assumed to follow the typical normal distribution.

This implies that a moving average model is a linear regression of the current observation of the time series against the random shocks of one or more prior observations.

To achieve greater flexibility in fitting of actual time series data, it is sometimes advantageous to include both autoregressive and moving average processes. This forms a general and useful class of time series models, known as the ARMA models.
Mathematically an ARMA(*p, q*) model is represented as

$$x_t = c + \omega_t + \sum_{i=1}^{p} \phi_i x_{t-i} + \sum_{j=1}^{q} \theta_j \omega_{t-j}$$

where *p* and *q* are autoregressive and moving average terms.

Usually ARMA models are manipulated using the lag operator (Cochrane, 1997, 2005; Hipel and McLeod, 2005) notation. The lag or backshift operator is defined as $Lx_t = x_{t-1}$. Polynomials of lag operator or lag polynomials are used to represent ARMA models as follows (Cochrane, 1997, 2005):

AR(*p*) model: $\omega_t = \phi(L)x_t,$

MA(*q*) model: $x_t = \theta(L)\omega_t,$

ARMA(*p,q*) model: $\phi(L)x_t = \theta(L)\omega_t,$

where $\phi(L) = 1 - \sum_{i=1}^{p} \phi_i L^i$ and $\theta(L) = 1 + \sum_{j=1}^{q} \theta_j L_j.$

The zeros of $\phi(L)$ must lie outside the unit circle for stationarity of the AR(*p*) process, and for invertibility of the MA(*q*) process the zeros of $\theta(L)$ must also lie outside the unit circle. This condition is known as the *Invertibility Condition* for an MA process.

2.3.9. Stationarity Analysis

Hipel and McLeod (2005) have shown that an important property of AR(*p*) process is invertibility, i.e. an AR(*p*) process can always be written in terms of an MA(∞) process. If AR(*p*) process is represented as $\omega_t = \phi(L)x_t$, then $\phi(L) = 0$ is known as the characteristic equation for the process. Box and Jenkins (1970) that a necessary and sufficient condition for the AR(*p*) process to be stationary is that all the roots of the characteristic equation must fall outside the unit circle. Hipel and McLeod (1994) also presented another simple algorithm for determining stationarity of an AR process. For

example as shown elsewhere the AR(1) model $x_t = c + \phi_1 x_{t-1} + \omega_t$ is stationary when $|\phi_1| < 1$, with a constant mean $\mu = \frac{c}{1-\phi_1}$ and constant variance $\gamma_0 = \frac{\sigma^2}{1-{\phi_1}^2}$.

An MA(q) process is always stationary, irrespective of the values the MA parameters Hipel and McLeod (1994). The conditions regarding stationarity and invertibility of AR and MA processes also hold for an ARMA process. An ARMA(p, q) process is stationary if all the roots of the characteristic equation $\phi(L) = 0$ lie outside the unit circle. Similarly, if all the roots of the lag equation $\theta(L) = 0$ lie outside the unit circle, then the ARMA(p, q) process is invertible and can be expressed as a pure AR process.

2.3.10. Autoregressive Integrated Moving Average (ARIMA) Models

In practice, many time series data exhibits non-stationary behaviour. Time series, which contain trend and seasonal patterns, are also non-stationary in nature (Faraway and Chatfield, 1998). Generally, ARMA models can be used for only stationary time series data. Thus ARMA models are inadequate to properly describe non-stationary time series, which are frequently encountered in practice. For this reason a generalisation of ARMA models which incorporates a wide class of non stationary time series as well is proposed (Box and Jenkins, 1970; Hipel and McLeod, 1994).

The integrated ARMA, or ARIMA, model is a broadening of the class of ARMA models to include differencing. The simplest example of a non stationary process which reduces to a stationary one after differencing is random walk. In ARIMA models, a non-stationary time series is made stationary by applying finite differencing of the data points. According to Shumway and Stoffer (2011), a process x_t is said to be ARIMA(p,d,q) if $\nabla^d x_t = (1-L)^d x_t$ is ARMA(p,q). In general, the model is written as

$$\phi(L)(1-L)^d x_t = \theta(L)\omega_t$$

where $\omega_t \sim WN(0, \sigma^2)$, WN indicating white noise. If $E(\nabla^d x_t) = \mu$, we write the model as

$$\phi(L)(1-L)^d x_t = \delta + \theta(L)\omega_t,$$

where $\delta = \mu\left(1 - \phi_1 - \cdots - \phi_p\right)$.

The integration parameter d is a nonnegative integer. When $d = 0$, ARIMA(p,d,q) ≡ARMA (p,q). An ARIMA(p,0,0) is nothing but the AR(p) model and ARIMA(0,0,q) is the MA(q) model. ARIMA(0,1,0), i.e. $x_t = x_{t-1} + \omega_t$ is a special one and known as the *random walk* model (Cochrane, 1997, 2005).

A useful generalization of ARIMA models is the Autoregressive Fractionally Integrated Moving Average (ARFIMA) model, which allows non-integer values of the differencing parameter *d*. ARFIMA has useful application in modelling time series with long memory (Galbraith and Zinde-Walsh, 2001). In this model the expansion of the term$(1- L)d$ is to be done by using the general binomial theorem. Various contributions have been made by researchers towards the estimation of the general ARFIMA parameters.

2.3.11. Seasonal Autoregressive Integrated Moving Average (SARIMA) Models

In this section, we introduce several modifications made to the ARIMA model to account for seasonal and non stationary behaviour. ARIMA models are used for non-seasonal non-stationary data. Box and Jenkins (1970,1976) have generalised this model to deal with seasonality. Their proposed model is known as the Seasonal ARIMA (SARIMA) model. In this model seasonal differencing of appropriate order is used to remove non-stationarity from the series. The fundamental fact about seasonal time series with period *S* is that observations, which are *S* intervals apart, are similar. Often, the dependence on the past tends to occur most strongly at multiples of some underlying seasonal lag S. Box and Jenkins (1970, 1976) proposed further that a seasonal series of period *S* could be modelled by

$$\Phi_P(L^S)\phi_p(L)\nabla_s^D\nabla^d x_t = \Theta_Q(L^S)\theta_q(L)\omega_t \qquad 2.1$$

where ω_t is the usual Gaussian white noise process. The general model in Equation (2.1) is denoted as SARIMA(*p,d,q*) x $(P,D,Q)_S$ and is called a *multiplicative seasonal autoregressive integrated moving average model*. For monthly time series $S = 12$ and for quarterly time series $S = 4$. The ordinary autoregressive and moving average components are represented by polynomials $\phi_p(L)$ and $\theta_q(L)$ of orders p and q respectively, and the seasonal autoregressive and moving average components by $\Phi_P(L^S)$ and $\Theta_Q(L^S)$ of orders P and Q and ordinary and seasonal difference operators by $\nabla^d = (1-L)^d$ and $\nabla_S^D = (1-L^S)^D$. For estimation of parameters, iterative least squares method is used.

2.3.12. Selection with the HK-algorithm

Hyndman and Khandakar (2008) developed the Hyndman-Khandakar (HK) algorithm and can be applied in R with the function auto.arima in the forecast package. They suggest an iterative time-saving procedure where the model with the smallest value of some information criterions AIC, AICc or BIC will be found much faster, since it is now found without comparing every possible model.

To derive these information criterions the first thing that is needed is the likelihood function, $L(\tilde{\varphi})$, where $\tilde{\varphi}$ is the maximum likelihood estimates of the parameters for the SARIMA with $k = p + q + P + Q + 1$ parameters and sample size *n*. The criterions are then derived by the following equations

$$AIC = 2k - 2\log[\, L(\tilde{\varphi})] \quad or \quad 2k + n\log\left(\frac{RSS}{n}\right)$$

$$AICc = AIC + \frac{2k(k+1)}{n-k+1}$$

$$BIC = -2\log[L(\tilde{\varphi})] + k\log(n) \quad or \quad \log(\sigma_e^2) + \frac{k}{n}\log(n)$$

where

k: is the number of parameters in the statistical model, (p+q+P+Q+1).

L: is the maximized value of the likelihood function for the estimated model.

RSS: is the residual sum of squares of the estimated model.

n : is the number of observation, or equivalently, the sample size.

σ_e^2 : is the error variance.

The AICc is a modification of the AIC by Hurvich and Tsai (1989) and it is AIC with a second order correction for small sample sizes. Burnham & Anderson (1998) insist that since AICc converges to AIC as *n* gets large, AICc should be employed regardless of the sample size. The HK-algorithm then performs an iterative procedure to select the model that minimizes the value of each criterion.

2.3.13. Multivariate ARIMA model

A multivariate time series is a combination of multiple univariate time series; simply called vector ARIMA (VARIMA) model involves a multivariate generalization of the univariate ARIMA model. Since VARIMA models can accommodate assumptions on exogeneity and on contemporaneous relationships, they offered new challenges to forecasters and policy makers. Work in this area started in the 1960s with population characteristics of VARMA processes by Quenouile (1957, 1968). Today, VARIMA models investigate the relationship between exogenous series and endogenous series where a dynamic system may exist i.e. in which a variation in the input series is utilised to explain a variation in the output series. Transfer function models are used to assess this relationship with input series and response series cross-correlated by way of a transfer function (TF). The exogenous variables can include continuous variables or dummy indicators highlighting the presence of an intervention or a stochastic series, which drives the response series. These types of

models are used to test explanatory relationships between time-dependent processes that are hypothesized to exist (Yaffee et al. 2000).

2.3.14. Transfer function

The dynamic or linear transfer function model can be useful (Pankratz, 1991) in overcoming possible problems of omitted time-lagged inputs terms, autocorrelation in the disturbance series, and common correlation patterns among the input and output series that yield spurious correlations. Notwithstanding, the identification of transfer function models can be difficult when there is more than one input variable. Edlund (1984) presented a two-step method for identification of the impulse response function when a number of different input variables are correlated. Using principal component analysis, a parsimonious representation of a transfer function model was suggested by del Moral & Valderrama (1997). Krishnamurthi et al. (1989) showed how more accurate estimates of the impact of interventions in transfer function models can be obtained by using a control variable.

A dynamic regression model, a term applied by Pankratz (1991) and used by Makridakis (1998), uses explanatory variables to forecast the dependent variable, but it still allows one to include the elements of ARIMA to model any patterns that cannot be accounted for by the explanatory variables. According to Makridakis (1998), they differ from multivariate autoregressive models in that the explanatory variables are leading indicators and are not affected by the dependent variable.

A dynamic regression model for one explanatory variable X can be written in two general forms as described in Makridakis (1998), but in the simpler form the forecast variable Y_t takes the form

$$Y_t = \alpha + \frac{\omega(L)}{\delta(L)} X_{t-b} + N_t$$

where X_t is the explanatory variable, where

$$\omega(L) = \omega_0 - \sum_{i=1}^{s} \omega_i L^i \quad \text{and} \quad \delta(L) = 1 - \sum_{j=1}^{r} \delta_j L^j$$

in terms of the backward shift operator (e.g. $LY_t = Y_{t-1}$) and where N_t is the combined effects of all other factors (i.e. noise, modelled as an ARIMA process). This formula extends naturally to several explanatory variables. In order to calibrate the model for one explanatory variable X , it is necessary to determine the values of r, s and b, as well as the values of $p, d,$ and p for the ARIMA(p, d, q) model for N_t . There are various methods for doing this. The method used in this study was suggested by Pankratz (1991) and

Makridakis (1998) and is referred to as the *Linear Transfer Function* (LTF) identification method.

2.3.15. Spectral Analysis

This is sometimes known as harmonic analysis or the frequency approach to time series analysis. Spectral analysis is therefore concerned with estimating the unknown spectrum of the process from the data and with quantifying the relative importance of different frequency bands to the variance of the process. The spectrum being estimated in a sense is not really the spectrum of the observed series, but the spectrum of the unknown infinitely long series from which the observed series is assumed to have come. Various methods have been developed to estimate the spectrum from an observed time series. For an overview and comparisons of different methods, see Percival and Walden (1993), Chatfield (2004), and Bloomfield (1976).

Two basic approaches to time series analysis are associated with the time domain or the spectral domain. The spectral domain approach is motivated by the observation that the most regular, and hence predictable, behaviour of a time series is to be periodic. This approach then proceeds to determine the periodic components embedded in the time series by computing the associated periods, amplitudes, and phases, in this order. The classical implementation of the spectral domain approach is based on the Bochner-Khinchin-Wiener theorem (Box and Jenkins, 1970), which states that the lag autocorrelation function of a time series and its spectral density are Fourier transforms of each other.

2.3.16. State Space Models

At the start of the 1980s, state space models were only beginning to be used by statisticians for forecasting time series, although the ideas had been present in the engineering literature since Kalman's (1960) ground-breaking work. State space models provide a unifying framework in which any linear time series model can be written. The key forecasting contribution of Kalman (1960) was to give a recursive algorithm (known as the Kalman filter) for computing forecast.

A particular class of state space models, known as "dynamic linear models" (DLM), was introduced by Harrison & Stevens (1976), who also proposed a Bayesian approach to estimation. Harvey (2006) provides a comprehensive review and introduction to this class of models including continuous-time and non-Gaussian variations.

Amongst this research on state space models, Kalman filtering, and discrete/continuous time structural models, the books by Harvey (1989), West & Harrison (1989, 1997) and Durbin & Koopman (2001) have had a substantial impact on the time series literature.

2.4. Dynamic Regression Models

Dynamic models have long been used in econometrics, agricultural econometrics and capital appropriations & expenditures. A class of dynamic models are the distributed lag models. Distributed lag models are useful because they allow a dependent variable to depend on past values of an explanatory variable at various lags. Therefore, decision makers or action planners can take into account the past or lagged values of the policy variables. This can be achieved through the use of many different models discussed in the literature that deal with this kind of situation.

Classical regression techniques are not designed to cope with variables that are non-stationary as they exhibit upward and downward trends over time. If explanatory variables exhibit such trends then classical assumptions will not simply work. In such instances, normal large-sample statistics theory is no longer valid and standard classical inferential procedures can no longer be applied.

In the late 1940s, Cochrane and Orcutt (1949) developed applications of least squares regression to relationships containing autocorrelated error terms. This was followed by Prais-Winsten (1954) and Hildreth-Lu (1960). During the same period, more efficient methods of estimation using distributed lag models were proposed. The distributed lag models received greater attention in the 1950s, when Koyck (1954), Cagan (1956), and Nerlove (1958b) suggested using an infinite lag distribution with geometrically declining weights for the parameters. For a thorough discussion of the Koyck model, see Nerlove (1958a). Additionally, there are several other models for reducing the number of parameters in a distributed lag model. Kmenta (1986), gave an overview of some of the most important distributed lag models such as the Pascal lag, the gamma lag, the LaGuerre lag and the Shiller lag models. For example, Pascal lag model is an infinite distributed lag model which is a flexible instrument for capturing the dynamic adjustment in most time series. Thomas (1997) clearly stated that for technological reasons, psychological factors and for imperfect information, distributed lag models should be used. However, Maeshiro (1996) and Thomas (1997) have pointed out that OLS estimation of the Koyck model gives inconsistent and biased estimators even if the sample size is increased indefinitely because

the equation involves lagged dependent variable and the errors are serially correlated. Almon (1965) developed polynomial distributed lags to approximate inverted U-shaped or even more complicated lag distributions that have a finite rather than an infinite maximum lag. Almon suggested that the immediate impact might well be less than the impact after several periods. After reaching its maximum, the policy effect diminishes for the remainder of the finite lag.

Bentzen and Engsted (2001) used the autoregressive distributed lag model in estimating the energy demand relationship. Hans and van Oest (2004) have also used distributed lag models to find the relationship between sales and advertising. More importantly, Huang et al. (2004) used Bayesian hierarchical distributed lag models in epidemiology for summer ozone exposure and cardio-respiratory mortality. Welty and Zeger (2005) used distributed lag models in environmental areas. Heaton and Peng (2013) investigated the effect of heat on mortality through the use of high degree distributed lag models. Schwartz et al. (1996) had already recommended that epidemiologists need to pay more attention to modelling distributed lags. For example, if we assume to use polynomial distributed lag models then we have three main issues - optimal lag length, order of the polynomial (Maddala, 1977; Hendry et al, 1984; Thomas, 1997; Maddala, 2009), and the difficulty in capturing any long-tailed distributions (Maddala, 2009). If any of these problems appear, the model may suffer from autocorrelation, heteroskedasticity, non-normality, incorrect functional form as well as the loss of degree of freedom among others. For more information and applications of distributed lag models you can see Cooper (1972), Shiller (1973), Fomby et al. (1984), Thomas (1997), Jeffrey and Wooldridge (2003), and Asteriou and Hall (2011).

2.5. Age-period-cohort (APC) Models

Age–period–cohort (APC) models have long been used in demography and medical statistics to describe the rate of mortality or incidence of a disease as a function of both age and period. Classically, APC models fit the effects of age, period, and cohort as factors (Hobcraft, Menken, and Preston 1982; Robertson, Gandini, and Boyle 1999).

Unfortunately, the use of these models is not straightforward as they suffer from an identifiability problem due to the exact linear relationship between age, period and cohort (Holford, 1983). This leads to a major challenge in analyzing APC models, a problem that has been widely addressed by statisticians, demographers and epidemiologists. The date of birth can be calculated directly from the age at diagnosis and the date of diagnosis (cohort

= period - age). If fitted directly in a generalized linear model (GLM) this leads to overparameterization and, consequently, the exclusion of one of the terms. It is therefore necessary to fit constraints to the model to extract identifiable answers for each of the parameters. This step is needed because each of the components of the model provides different insights into the trends of the disease over time. The multiple classification model which is the initial work developed by Mason et al. (1973) presented the general framework for cohort analysis. For example, in social and demographic research, Glenn (1976), Fienberg and Mason (1978, 1985), Hobcraft, et al. (1982), Wilmoth (1990), and O'Brien (2000) followed a number of methodological discussions to overcome the identifiability problems and to estimate the APC model. Also, in epidemiology and biostatistics, Osmond and Gardner (1982), Clayton and Schifflers (1987), Holford (1992), Tarone and Chu (1992), Robertson and Boyle (1998), Fu (2000), Knight and Fu (2000), Yang et al. (2004), Carstensen (2007), Rutherford et al. (2010), Rutherford et al. (2012), and Sasieni (2012) have proposed a number of solutions to solve the identifiability problems over the past 30 years.

Developments in APC methodology in biostatistics over the past three decades have stressed the use of estimable functions that do not respond to the selection of constraints on the parameters (Clayton and Schifflers 1987; Holford 1983, 1991, 1992; Robertson et al. 1999; Tarone and Chu 1992, 2000). For more details on estimable functions see, for example, Searle (1971). Fu et al. (2004) used this approach in the derivation of a new APC estimator called intrinsic estimator. This estimator is based on estimable functions and the singular value decomposition of matrices. Yang et al. (2004), on the other hand, used the conventional demographic approach of constrained generalized linear models estimator (Fienberg and Mason 1978, 1985; Mason and Smith 1985) and the intrinsic estimator method developed by Fu 2000; Knight and Fu 2000; Fu et al. 2004 to compare parameter estimates and model fit statistatics produced by two solutions to the identification problem in APC models. The two approaches to solving the model identification problem in APC models are described in detail and compared by Yang et al. (2004). Carstensen (2007) published an article advocating the use of an analysis that models age, period, and cohort as continuous variables through the use of spline functions. Carstensen implemented his method for age–period–cohort models in the Epi package for R. Rutherford et al. (2010, 2012) built on the work of Carstensen and explained how the method and the extensions have been made available in Stata. Sasieni (2012) fully explained and illustrated programs including postestimation functionality and flexibility to fit models not possible using

Stata's glm command as described by Rutherford et al. (2010). What distinguishes this article from a recent *Stata Journal* article on age–period–cohort models by Rutherford et al. (2010) is that the emphasis made by Sasieni (2012) is on extrapolating the model fit to make projections into the future.

Bayesian APC models are used more frequently in the last few years in epidemiology, demography, social & political behaviour and cancer research to predict cancer incidence and mortality rates (Baker and Bray 2005; Raifu and Arbyn 2009). The Bayesian APC model provides an effective way to cope with the identification problem inherent in the model and offer better predictions than the classical APC approaches. It has been found that Bayesian APC models do not pose any implementation problems when there are many zero counts or sparse data, whereas the classical APC models may lead to instable parameters estimates in this case (Raifu and Arbyn, 2009). Moreover, Bayesian APC models are recommended recently because they reduce the errors associated with functions of the parameters as much as possible by smoothing the effect of age, period and cohort (Cleries et al., 2010).

Bayesian APC approaches were proposed firstly by Berzuini et al., (1993), Berzuini and Clayton (1994), and Besag et al., (1995). To reduce variation of the model parameters, several methods have been proposed during the last 30 years, in such a way that the identification issue is avoided. For example, Nakamura (1986) used a first-order autoregressive approach whereas Berzuini and Clayton (1994) used a second-order random walk. Besag et al., (1995) proposed a sophisticated MCMC algorithm using reparameterisation and block sampling to fit a Bayesian APC model using the second-order random walk. Rue et al., (2009) proposed an alternative and fast method of inference which is an Integrated Nested Laplace Approximations (INLA). This is because improper priors can generate problems in making inference. Therefore, prior distributions should be carefully selected based on previous studies in the literature or on subjective prior beliefs.

Breslow & Clayton (1993), and Berzuini & Clayton (1994) use Bayesian APC to model breast cancer and lung cancer respectively. Besag et al., (1995) used Bayesian logistic regression to forecast prostate cancer in USA. Leonhard et al., (2001) used a generalized Bayesian APC model to predict lung cancer mortality in West Germany by 2010. Bray (2002) fitted a Bayesian APC model to predict incidence rates for Hodgkin's disease for males registered in Oxford. Cleries et al., (2006) used an autoregressive structure using Bayesian approach to predict breast cancer mortality rates in Spain. Raifu and Arbyn (2009) used a Bayesian log linear Poisson-regression model to assess the effects

of age, period and cohort. Stegmueller (2014) proposed a Bayesian dynamic hierarchical model with cohort and period effects modelled as random walk through time to model cancer in the USA.

2.6. Methods for Quantification of Incidence and Mortality

2.6.1. Methods and Techniques

The source of the following definitions and terminologies can be found in Cancer Atlas of Saudi Arabia (2011).

2.6.2. Rates

Rate expresses how often a disease (cancer) occurs in a given population over a given period of time.

2.6.3. Age-specific Rates

The age specific rates per year are obtained from the cancer registries. The all ages rate referred to as crude rate is defined as follows:

$$Crude\ Rate = \frac{Total\ registrations}{Total\ population} \times 100{,}000 \qquad 2.2$$

Calculation of age-specific rates for each age group can also be defined as follows:

$$AsR_k = \left(\frac{r_k}{p_k}\right) \times 100{,}000 \qquad 2.3$$

where AsR_k is the age-specific rate for age group k, r_k is the number of registrations in age group k, p_k is the people at risk in age group k and k is the group index for age groups 0-4, 5-9,..., 70-74, and 75+.

It is possible to calculate the age specific rates of lung cancer incidence separately for females and males, or for both genders combined. In order to make comparisons of incidence rates over time or between genders and geographical areas, age standardised rates are used to make unbiased comparisons.

2.6.4. Age-standardised Rates

Lung cancer incidence and mortality vary greatly with age. Therefore, to specify how many old or young people are in the population being looked at, we use age standardized rates in order to obtain unbiased comparisons of incidence or mortality rates between genders or regions over time. Thus, if lung cancer rates are not age standardised, a higher

rate in one country is likely to reflect the fact that it has a higher proportion of older people. This can be obtained through direct or indirect standardisation according to dos Santos Silva (1999). Throughout our research we use the direct stanadrdised rate.

Age standardized rates can be calculated directly by multiplying the age specific rate in each group in the populations by the corresponding number of people in a 'standard' population, usually the world standard population – see Appendix F of *Cancer Trends* (Quinn et al, 2001), and then summed to give the overall rate of lung cancer per 100,000 population. Thus,

$$ASR = \left\{\sum_k Z_K \; W_k\right\} / \sum_k Y_k \qquad 2.4$$

where ASR is the age standardised incidence or mortality rate, Z_k is the number of cases of lung cancer in age group k, W_k is the world standard population in age group k, Y_k is persons at risk and k = 0-4, 5-9,..., 70-75, and 75+.

To study cancer incidence directly, we should take into account the age as a major determinant of cancer incidence. It has been stressed in each consecutive volume that the most suitable comparisons of cancer risk are those made using the age specific rates directly. Most developed countries have taken a higher proportion of elderly people into account to make a comparable comparison between developed and developing countries because the elderly are expected to live longer in developed countries than in developing countries.

2.7. Prediction Methods

There are several methods used in predicting or forecasting cancer in general but APC models have been used widely to predict cancer incidence and mortality rates and are unique (Holford, 1985). In most well developed countries, they use the age-period cohort models, known as the APC model.

Many methods have been proposed for making projections from APC models. Good references for APC model projections are Clements et al. (2005), Elkum (2005), Bray and Moller (2006), and Rutherford et al. (2012). Bray et al. (2001), Cleries et al. (2009), Lee et al. (2011), and Mistry et al. (2011) to mention a few. Moller et al. (2003) compared fifteen of these methods using data from the Nordic countries. Sasieni and Adams (1999, 2000) used natural cubic splines in APC models for drawing inference on the impact of cervical screening on cervical cancer rates. Carstensen (2007) has written about their use more generally. Rutherford et al. (2010) has provided software in Stata for fitting APC models

using natural cubic splines. Quite apart from these methods, a good overview of techniques available to carry out APC model projections using natural cubic splines has been given by Rutherford et al. (2012) and Sasieni (2012). They concluded that multiplicative APC models tend to over-estimate future incidence and therefore linear projections need to be tempered or dampened when making long-term prediction. For that reason, they advocated the use of an APC model with a power link function together with a linear combination of age, period and cohort terms.

The main advantage of using the APC models is that they take into account the period, cohort, and age effects to forecast the future temporal trends of cancer rates. However, it has been advised by Clytton and Schifflers (1987a and 1987b) that reduction of the APC model to be either an age-period (AP) model or an age-cohort (AC) model whenever possible is better, only using the APC model when it provides a satisfactory fit.

2.8. Summary

To identify the potential areas of our research and to ensure that we have a full understanding of the problem, we have reviewed the literature to identify similar works, to compare previous findings and to suggest future work.

Time series analysis is frequently used in statistics, econometrics, mathematical finance and weather forecasting among many other fields to analyse time series data in order to extract meaningful statistics and other characteristics of the data. A variety of models have been proposed in literature to improve the accuracy and efficiency of time series modelling and forecasting. Generally, time series analysis falls into two main approaches: the time domain analysis; and the frequency domain analysis. There are many techniques available to analyse data within each domain. One common technique used in the time domain is the Box-Jenkins methodology, which can be used for univariate or multivariate analyses. Analysis in the frequency domain is often used for periodic and cyclical observations. Common techniques are spectral analysis, periodogram analysis, and harmonic analysis. Mathematically, frequency domain techniques use fewer computations than time domain techniques. Thus, for complex data, analysis in the frequency domain is most common. However, frequency analysis is more difficult to interpret, so time domain analysis is generally used outside of the sciences.

Distributed lag models (DLM) have been reported in the literature since the early 1930s. Distributed lag models are useful because they allow a dependent variable to depend on past values of an explanatory variable at various lags. The difficulty of using

these models requires choosing the optimal lag length, order of the polynomial and capturing any long-tailed distribution. In this case, if the model is not specified correctly, the model will suffer from autocorrelation, heteroskedasticity, non-normality, incorrect functional form and the loss of degrees of freedom. The models, their difficulties, and corrections have been explicitly explained.

Age-period-cohort (APC) models are the most popular tools used in cancer studies to describe the rate of incidence or mortality as a function of both subject age and period. However, the use of these models is known to suffer from an identifiability problem due to the exact linear relationship between age, period and cohort. New approaches have been developed for APC analysis to overcome the identification problem during the last 30 years. Overcoming the identification problem by forcing constraints on either the period or the cohort effects has been emphasized.

Our thesis will help the Saudi Arabian Ministry of Health to understand the rate of future lung cancer incidence and mortality and the overall effects of the population classes and budgeting costs needed for lung cancer in Saudi Arabia. We therefore expect that our thesis will produce an impact on Saudi Arabian health policy.

CHAPTER 3

DATA FOR THE RESEARCH PROJECT

3.1. Introduction

Lung cancer incidence and mortality data in Saudi Arabia between 1994 and 2009 were collected from Saudi Cancer Registry (SCR). These data include the date of diagnosis, gender, ethnicity, type of lung cancer, region, age at diagnosis, date of birth, and the status (dead, alive, or unknown).

The Central Department of Statistics & Information (CDS) in the Ministry of Planning provided data on person characteristics, such as age, gender, and ethnicity. Other data including estimated population of both Saudi and non-Saudis are tabulated ranging from Table F1 to Table F20 (see Appendix F). The estimated Saudi male population in 2009 was 9,216,449 accounting for 35.1% of the total population and is tabulated in Table F2, whereas the Saudi female population of about 8,855,113 accounts for 33.7% of the total population (see Table F6). The non-Saudi populations were 5,784,649 (22.0%) for males and 2,434,016 (9.3%) for females and are shown in Table F4 and Table F8 respectively.

3.2. Incidence Data

Figure 3.1 shows the number of cases of lung cancer incidence in KSA per year for Saudi males, non-Saudi males, Saudi females and non-Saudi females from 1994 to 2009. Data for Figure 3.1 are shown in Table F1, F3, F5 and F7. Notably, the reduction in numbers of cancers in some cases at increased age is due primarily to the reduction in the associated number of individuals at risk. In such situations, to accommodate for the number of individuals at risk, we focus on the incidence rate, which is equal to the number of events divided by the number at risk and multiplied by one hundred thousand. Therefore, we present the age-specific incidence rates per 100,000 population for Saudi males in Table 3.1, non-Saudi males in Table 3.2, Saudi females in Table 3.3 and non-Saudi females in Table 3.4. The data have been also presented separately for males and females in months from 1994-2009 as illustrated in Figure3.2 (see also Table F9 and Table F10).

The data of smoking prevalence were collected from the Department of Tobacco Control Program in the Ministry of Health. Figure 3.3 illustrates smoking populations in 10,000 per month for males (X_{1t}) and females (X_{0t}), respectively. The data for smoking populations are shown in Table F11 and Table F12. (Note: smoking population (X_t) = smoking prevalence (%) × population size).

SCR is located in King Faisal Specialist Hospital and Research Centre in Riyadh (KFSH & RC). In addition, five regional branches and four hospital-based offices were set up to ensure comprehensive data collection from all over the kingdom. They are National Guard Hospitals, Armed Forces Hospitals, Security Forces Hospitals, King Abdulaziz University Hospital, King Khalid University Hospital, Madinah Region, Southern Region, Eastern Region, Western Region, and Central region (KFSH & RC).

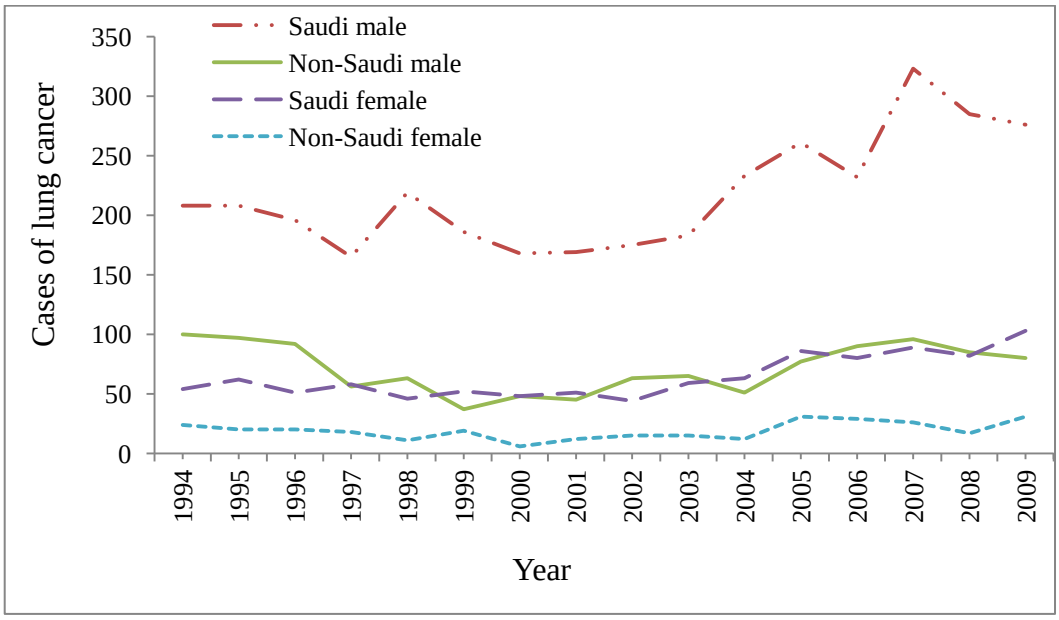


Figure 3.1: Number of cases of lung cancer per year by ethnicity and gender from 1994 to 2009.

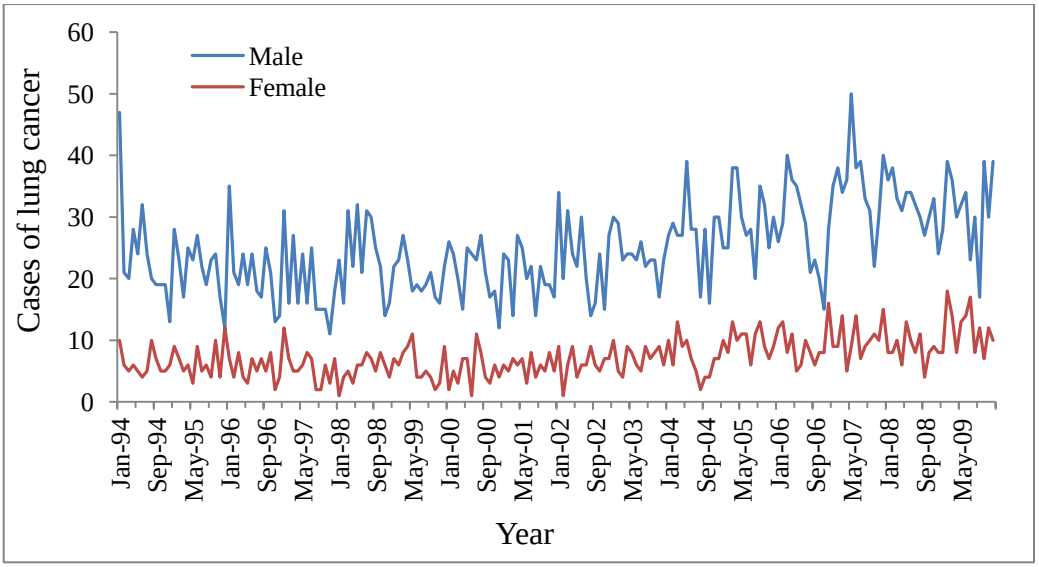


Figure 3.2: Number of cases of lung cancer per month in Saudi Arabia by gender from 1994 to 2009.

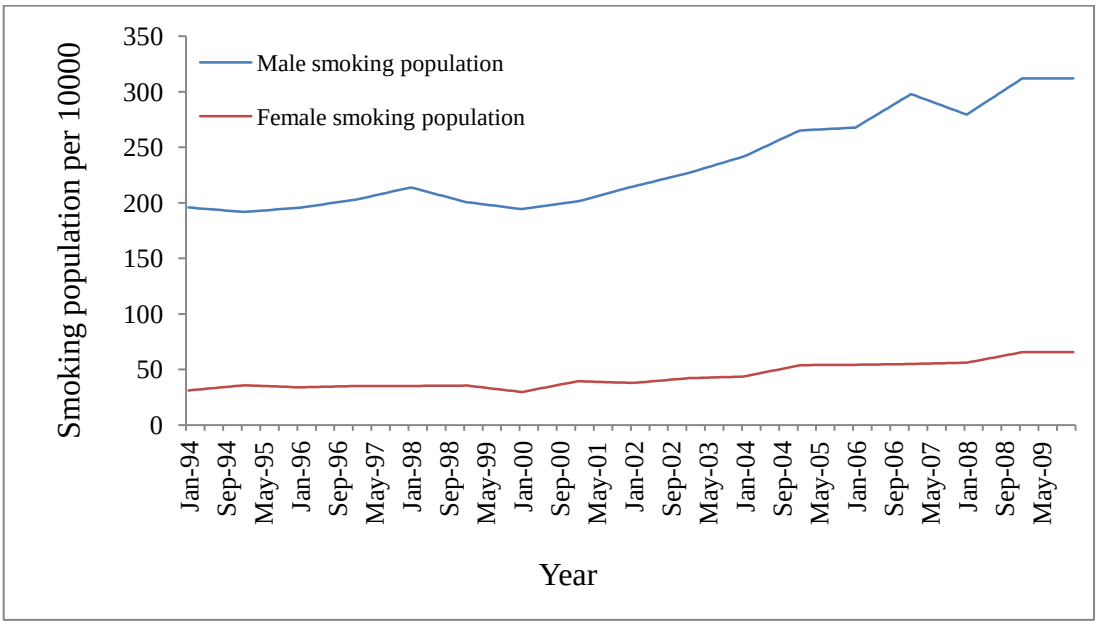


Figure 3.3: Smoking population in Saudi Arabia by gender from 1994 to 2009.

Table 3.1: Age-specific incidence rates of lung cancer per 100,000 population for Saudi males in KSA (1994-2009).

	Time period (1994-2009)															
Age	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09
0 - 4	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.1
4-9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
10-14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
15-19	0.0	0.1	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.0
20-24	0.2	0.2	0.0	0.0	0.2	0.0	0.1	0.1	0.0	0.0	0.1	0.0	0.1	0.2	0.0	0.1
25-29	0.2	0.2	0.0	0.4	0.4	0.2	0.0	0.2	0.0	0.1	0.0	0.1	0.1	0.3	0.3	0.3
30-34	0.3	0.8	1.4	0.5	1.0	1.0	0.5	0.4	0.4	0.5	0.7	0.5	0.3	0.2	0.4	0.6
35- 39	1.9	2.2	1.8	2.5	0.9	1.4	0.8	1.3	1.3	1.0	0.8	2.0	0.2	1.9	0.2	0.4
40-44	6.5	5.3	0.9	3.7	1.8	3.0	1.6	0.6	1.5	2.0	2.7	3.3	1.4	3.4	2.4	1.7
45-49	7.0	5.9	4.8	4.8	7.9	1.9	3.4	4.1	2.6	1.3	4.8	5.3	4.5	2.8	4.7	3.8
50- 54	18.2	14.7	15.9	12.7	10.7	10.5	4.3	9.7	6.4	8.3	11.7	6.2	5.6	9.5	12.4	9.7
55- 59	27.1	25.2	16.8	19.8	18.8	9.0	13.3	16.8	11.6	9.2	15.8	12.0	13.3	12.4	13.9	12.3
60- 64	327.3	62.7	33.6	19.7	22.3	20.6	29.5	26.4	38.9	22.8	35.4	32.3	21.3	32.7	26.5	22.8
65- 69	45.5	60.7	57.1	29.2	39.1	38.1	26.4	13.8	30.1	35.6	23.3	42.5	61.8	56.3	41.5	52.8
70- 74	46.8	50.8	45.9	30.0	56.3	32.2	31.9	29.8	28.7	37.3	40.3	58.2	51.8	52.8	39.1	44.1
75+	47.0	35.7	48.8	25.5	23.9	32.5	24.8	8.2	16.7	35.4	38.8	37.0	55.4	64.7	55.0	45.9

Table 3.2: Age-specific incidence rates of lung cancer per 100,000 population for non-Saudi males in KSA (1994-2009).

	Time period (1994-2009)															
Age	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09
0 - 4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5-9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10-14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15-19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
20-24	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25-29	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
30-34	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
35- 39	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
40-44	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
45-49	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
50- 54	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.1	0.1	0.1	0.1
55- 59	0.3	0.4	0.3	0.2	0.1	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.2	0.1	0.1
60- 64	0.6	0.6	0.7	0.3	0.4	0.3	0.1	0.4	0.1	0.1	0.2	0.1	0.4	0.4	0.3	0.2
65- 69	1.0	1.0	1.3	0.4	0.4	0.1	0.5	0.4	0.7	0.9	0.4	0.5	0.7	0.5	0.6	0.4
70- 74	0.8	0.8	1.2	0.5	0.2	0.4	0.7	0.1	0.4	0.3	0.2	0.5	0.6	0.5	0.7	0.5
75+	1.2	1.2	0.8	0.8	0.2	0.4	0.3	0.0	0.1	0.0	0.4	0.8	0.4	0.5	0.2	0.4

Table 3.3: Age-specific incidence rates of lung cancer per 100,000 population for Saudi females in KSA (1994-2009).

	Year of diagnosis (1994-2009)															
Age	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09
0 - 4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5-9	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1
10-14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15-19	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.0	0.0	0.1
20-24	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.4	0.0	0.0	0.0	0.2	0.0
25-29	0.2	0.0	0.0	0.4	0.2	0.0	0.0	0.2	0.2	0.3	0.1	0.1	0.5	0.4	0.1	0.4
30-34	0.3	0.3	1.3	0.8	0.0	0.0	0.0	0.0	0.4	0.5	0.0	0.3	0.2	0.0	0.3	0.3
35- 39	1.3	1.0	0.3	0.6	0.9	0.5	0.0	0.2	0.5	0.4	0.2	0.6	0.2	0.4	0.2	0.4
40-44	1.9	0.9	0.5	1.8	0.9	1.1	0.6	0.6	1.2	1.1	0.3	2.6	1.4	0.9	0.7	1.6
45-49	2.7	3.3	2.2	1.1	1.0	1.4	0.9	2.0	1.1	0.7	1.8	3.9	2.4	3.0	2.3	2.3
50- 54	4.2	2.8	4.2	4.6	2.7	3.1	4.5	1.0	1.4	4.1	2.5	3.9	2.0	3.7	1.5	7.3
55- 59	3.2	6.9	4.4	4.0	6.1	6.2	2.7	3.2	3.6	4.0	6.5	4.5	3.4	3.1	5.1	3.6
60- 64	5.1	6.0	3.4	6.9	5.3	8.3	3.9	7.2	3.7	6.9	9.7	9.4	9.6	4.8	9.9	6.0
65- 69	16.9	11.7	4.9	8.5	10.0	4.1	10.3	8.5	6.6	6.1	6.9	2.9	14.7	17.6	9.9	17.1
70- 74	4.5	7.0	13.3	7.1	4.6	4.3	8.0	7.9	4.2	6.0	7.1	12.6	15.3	10.3	13.8	7.5
75+	9.5	13.7	6.6	11.2	12.4	15.1	11.5	12.6	8.2	13.2	10.8	16.7	14.6	17.2	11.8	15.1

Table 3.4: Age-specific incidence rates of lung cancer per 100,000 population for non-Saudi females in KSA (1994-2009).

	Time period (1994-2009)															
Age	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09
0 - 4	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5-9	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10-14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15-19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
20-24	0.0	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0
25-29	0.7	0.0	0.0	0.0	1.1	0.0	0.0	0.0	0.7	0.4	0.4	0.0	0.0	0.0	0.4	0.0
30-34	2.1	1.2	0.8	0.0	0.7	0.0	0.0	0.0	1.4	1.1	0.4	0.4	1.7	0.0	0.3	0.5
35- 39	0.0	1.2	1.8	0.6	0.6	1.1	0.5	0.0	0.0	0.0	0.5	1.5	0.4	0.8	0.0	0.0
40-44	3.3	1.1	1.1	2.2	3.4	0.0	0.0	0.0	2.1	2.5	0.0	2.3	4.5	0.7	0.7	2.5
45-49	2.0	4.2	4.3	4.4	0.0	8.2	1.9	3.6	3.0	2.8	1.4	0.0	2.7	6.8	1.3	3.3
50- 54	12.0	7.7	7.4	11.1	0.0	10.3	0.0	3.2	8.6	2.3	6.8	6.5	2.2	10.6	6.3	7.0
55- 59	30.0	45.5	27.3	25.0	8.3	13.3	16.7	0.0	5.0	14.3	4.5	21.7	11.5	15.4	3.7	9.4
60- 64	16.7	25.0	45.5	36.4	0.0	27.3	9.1	18.2	0.0	0.0	6.7	6.3	35.7	13.3	6.7	16.7
65- 69	50.0	33.3	40.0	40.0	20.0	33.3	0.0	33.3	14.3	12.5	12.5	75.0	33.3	11.1	33.3	54.5
70- 74	20.0	0.0	0.0	0.0	20.0	14.3	0.0	22.2	0.0	16.7	16.7	14.3	37.5	44.4	11.1	30.0
75+	16.7	0.0	0.0	0.0	0.0	0.0	0.0	16.7	14.3	0.0	0.0	62.5	0.0	25.0	50.0	30.0

Table 3.5 shows the overall age-specific incidence rate of lung cancer in one-year intervals from 1994 to 2009 and five-year age groups from 0-4 years to 75+ in Saudi Arabia. The rates show an increasing incidence of lung cancer with increasing age to 65-69 in all sixteen time periods. Among older age groups, there is a 50 per cent increase in incidence rates from 1994-2009 but in the age groups under 35 the increases are based on very limited absolute numbers. Such a pattern points to an interaction between age group and time period. This may be of significance for the aetiology or may reflect an increase in the completeness of coverage for registration of incident cases, or components of both. Our initial task of fitting the APC models is to estimate the effects of each of these three factors on the rates. Figure 3.4 shows the average incidence rate of lung cancer per 100,000 for the 16 age groups from 1994 to 2009.

Table 3.5: Overall age-specific incidence rates per 100,000 of lung cancer population in KSA 1994-2009.

	Year of diagnosis (1994-2009)															
Age	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09
0 - 4	0.1	0.0	0.0	0.0	0.2	0.1	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.1
5-9	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1
10-14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
15-19	0.0	0.2	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.3	0.1	0.1	0.1	0.0	0.1
20-24	0.2	0.4	0.0	0.1	0.2	0.0	0.1	0.2	0.0	0.0	0.5	0.1	0.1	0.3	0.2	0.2
25-29	0.4	0.1	0.0	0.5	0.6	0.1	0.4	0.2	0.5	0.6	0.3	0.2	0.5	0.5	0.3	0.4
30-34	1.1	0.9	1.8	0.7	0.8	0.4	0.3	0.2	1.1	1.1	0.5	0.6	1.0	0.1	0.7	0.8
35- 39	1.9	2.3	1.7	1.6	1.9	1.6	0.8	1.1	1.1	1.0	0.8	2.1	0.5	1.6	0.6	0.7
40-44	4.9	3.6	2.4	4.3	4.1	3.1	2.1	1.2	2.8	3.0	1.6	5.1	3.5	2.9	2.3	3.0
45-49	9.0	9.6	7.5	5.3	6.9	4.6	3.7	5.9	3.9	2.9	4.9	7.5	6.0	7.0	5.2	4.9
50- 54	18.5	15.5	17.6	16.7	12.1	12.6	9.6	9.5	9.1	11.2	11.7	9.8	8.2	13.0	12.5	14.8
55- 59	35.0	38.8	25.8	26.4	24.1	14.5	15.6	17.4	16.0	15.9	19.7	20.1	16.6	18.8	17.7	14.6
60- 64	71.1	70.2	46.0	30.3	30.1	31.4	30.4	37.4	34.7	25.9	41.9	37.2	37.2	39.7	37.7	28.7
65- 69	72.1	79.3	73.6	41.2	50.0	41.7	38.5	26.5	43.6	49.3	33.2	51.0	80.1	72.6	55.0	71.0
70- 74	55.7	59.9	64.6	38.0	59.1	38.3	41.9	38.1	33.0	43.2	45.6	69.8	70.8	66.1	56.3	54.8
75+	61.2	53.9	57.2	39.2	35.1	47.3	35.7	20.6	25.0	45.1	49.2	60.2	66.6	81.5	65.5	62.1

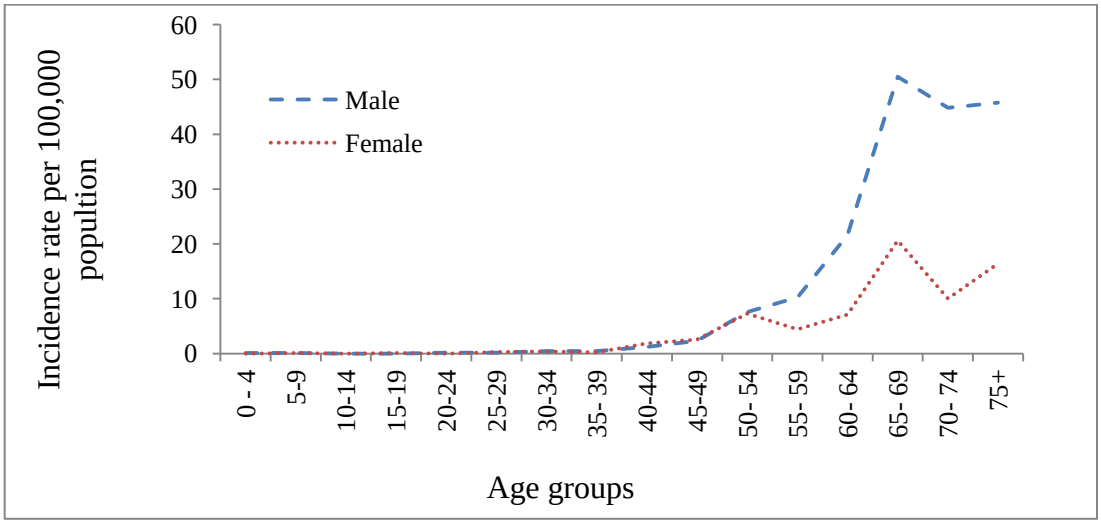


Figure 3.4: Average incidence rate of lung cancer per 100,000 for the 16 age groups from 1994 to 2009.

Table 3.6 shows the total cases of lung cancer by region, price of imported tobacco in millions of dollars and consumption of tobacco in thousands of tons from 1994 to 2009.

Notice that the 13 administrative regions in Saudi Arabia are divided into five regions in our study in order to obtain an overall picture of the future lung cancer burden in KSA. These five regions are presented in Figure 3.5. Thus, the northern region includes Tabuk, Hail, Jouf and Northern Border cities. The southern region includes Asir, Baha, Najran and Jazan cites. The western region includes Makkah and Madinah cites whereas, the central

region includes Riyadh and Qassim cites. The eastern region includes the whole Eastern province.

Our aim of including various covariates in this thesis is to establish a more realistic model of the relationship between some environmental lifestyles and lung cancer incidence in all age groups for males and females across Saudi Arabia. Initially, our desire not to include as many explanatory variables including smoking level and alcohol consumption were due to inaccessibility of data. However, a set of variables, namely, gender, race, age, consumption of tobacco per 1000 tons, smoking prevalence by gender, and five regions of Saudi Arabia were mentioned.

There were some problems with the data. For example, the Ministry of Health encountered problems at the beginning during the diagnostic, monitoring, treatment period and collection of the data. These problems were due to the untrained staff at the Saudi Cancer Registry at the time, influx of foreign nationals or immigration caused by the Gulf war and probably the lack of modern diagnostic techniques using technological resources. Another reason could be due to poor case ascertainment and certification at older ages (Saudi Cancer Registry, 2009).

Table 3.6: Total cases of lung cancer by region, price of imported tobacco in millions of dollars and consumption of tobacco in 1000 tons from 1994 to 2009.

Year	Consumption in (1000 tons)	Price (millions)	Northern Cases	Southern Cases	Western Cases	Central Cases	Eastern Cases
1994	9	401	15	22	115	93	83
1995	22	844	18	27	127	77	70
1996	29	633	19	21	144	53	62
1997	39	1353	16	17	98	100	53
1998	39	1353	12	15	128	90	61
1999	37	1300	19	21	115	84	50
2000	36.5	1320	21	13	95	91	72
2001	37.7	1450	23	14	78	85	83
2002	38.8	1460	23	20	107	104	79
2003	39.2	1500	13	30	141	83	76
2004	43.6	1600	23	18	169	79	85
2005	44.2	1700	10	21	201	124	74
2006	46	1750	25	23	189	125	77
2007	43	2058	44	31	204	125	115
2008	47	2264	34	21	184	121	107
2009	52	2491	26	33	187	144	114

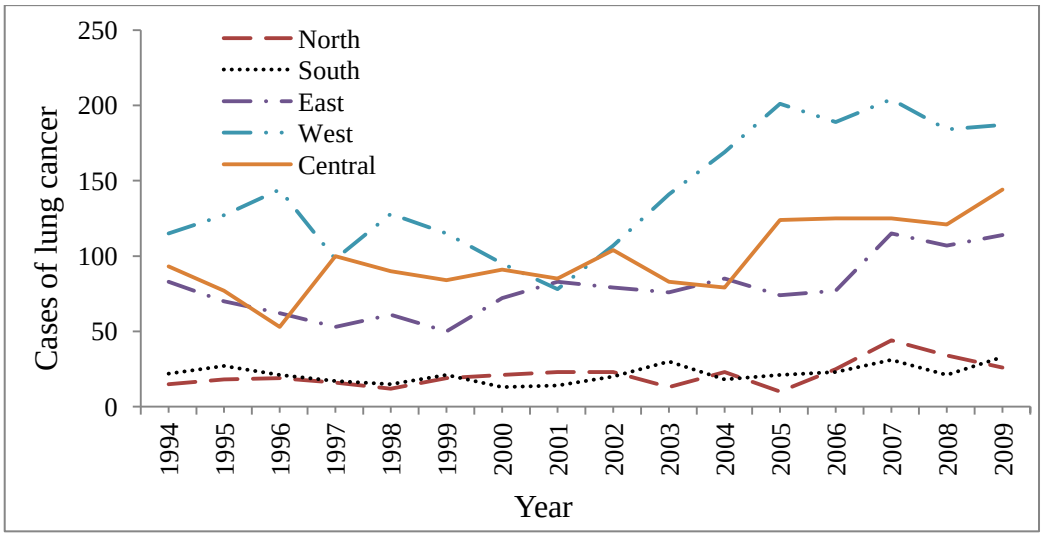


Figure 3.5: Number of cases of lung cancer per year by regions in KSA from 1994 to 2009.

3.3. Mortality Data

The cases of lung cancer mortality per year are presented in Figure 3.6 for males and females separately from 1994 to 2009. Data for Figure 3.6 are shown in Tables F15 and F16. In addition, the mortality data in months are presented separately for males and females from 1994 to 2009 as shown in Figure 3.7. We have also presented the mortality data in Tables F17 and F18. The age-specific mortality rates of lung cancer per 100,000 population for both males and females combined are presented in Table 3.7. We arrange the data in one-year intervals from 1994 to 2009 and five-year age groups from 25-29 years to 75+ years.

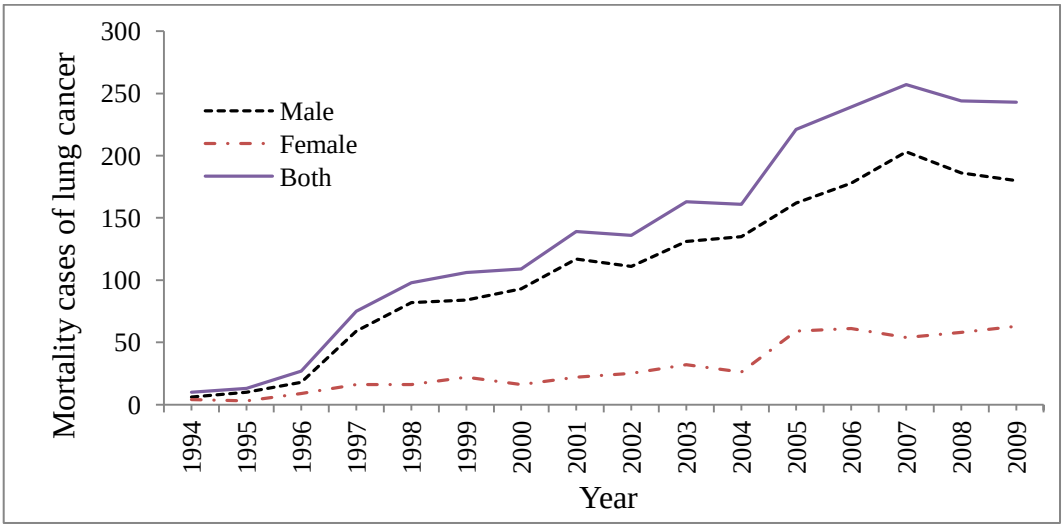


Figure 3.6: Number of cases of lung cancer mortality per year by gender from 1994 to 2009.

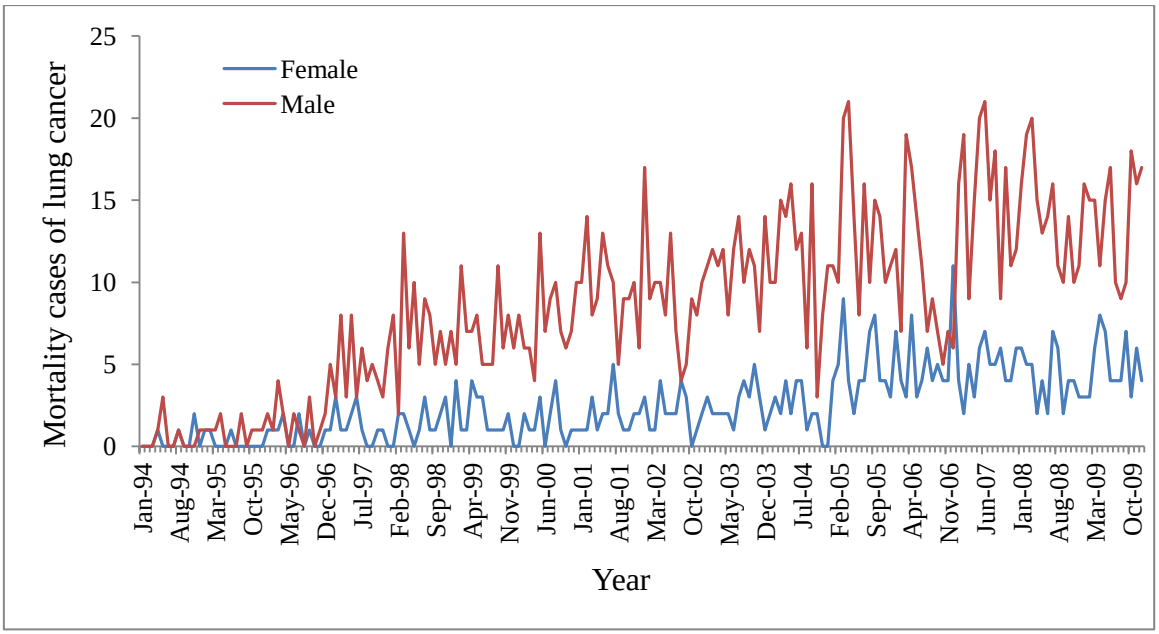


Figure 3.7: Number of cases of lung cancer mortality per month in Saudi Arabia by gender from 1994 to 2009.

Table 3.7: Age-specific mortality rates per 100,000 of lung cancer for population in KSA 1994-2009.

	Year of diagnosis (1994-2009)															
Age	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09
25-29	0.0	0.1	0.0	0.1	0.1	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.1	0.0
30-34	0.1	0.0	0.1	0.2	0.0	0.2	0.1	0.1	0.1	0.0	0.1	0.1	0.3	0.0	0.1	0.1
35-39	0.1	0.1	0.1	0.1	0.3	0.3	0.2	0.2	0.3	0.5	0.1	0.4	0.0	0.2	0.0	0.1
40-44	0.0	0.2	0.0	0.1	0.3	0.4	0.4	0.1	0.4	0.7	0.6	0.7	0.6	0.6	0.7	0.5
45-49	0.2	0.5	0.3	0.3	1.1	0.6	0.8	1.2	0.8	1.6	0.6	1.4	1.8	1.8	1.2	1.2
50-54	0.0	0.0	0.7	3.2	1.2	3.3	1.7	3.6	1.3	2.2	2.9	2.2	1.8	2.8	3.5	3.4
55-59	0.7	0.7	1.3	5.5	5.0	2.0	2.5	6.0	3.5	3.8	3.2	4.3	3.9	4.2	4.6	4.3
60-64	0.7	0.7	1.4	2.7	6.5	4.1	5.9	8.1	10.6	9.1	10.1	10.2	10.8	11.5	10.9	7.9
65-69	1.4	0.7	3.9	4.4	8.0	7.5	9.5	6.5	7.3	11.2	8.3	15.7	20.0	17.2	9.0	18.6
70-74	0.0	0.7	2.7	2.0	9.0	7.0	8.1	11.8	11.1	11.5	13.5	18.0	25.9	21.0	17.8	13.7
75+	0.6	0.0	0.5	7.7	6.1	12.3	9.1	9.5	9.8	10.5	15.1	20.3	18.9	21.0	22.7	19.6

3.4. Population Forecast by 2020

The estimated male population in 2009 was 15,010,101 accounting for 57% of the total population, and female population was 11,325,130 accounting for 43.0% of the total population (see Figure 3.8). The Department of Economic and Social Affairs at the United Nations in 2012 made the forecasts of population growth between 2010 and 2020. It assumed that males would experience the largest proportional increase by 16.2% as shown in Table F19, whereas females were estimated to increase by 6.4% in 2020, which is illustrated in Table F20. The reason for this is that female forecasts assume a relatively low birth rate and low net immigration. The total population estimated by 2020 is 32,340,000. Its structures vary by the age distribution for males and females separately as in Figures 3.9 and 3.10. Figure 3.11 shows the age distribution of the world standard population in 2009.

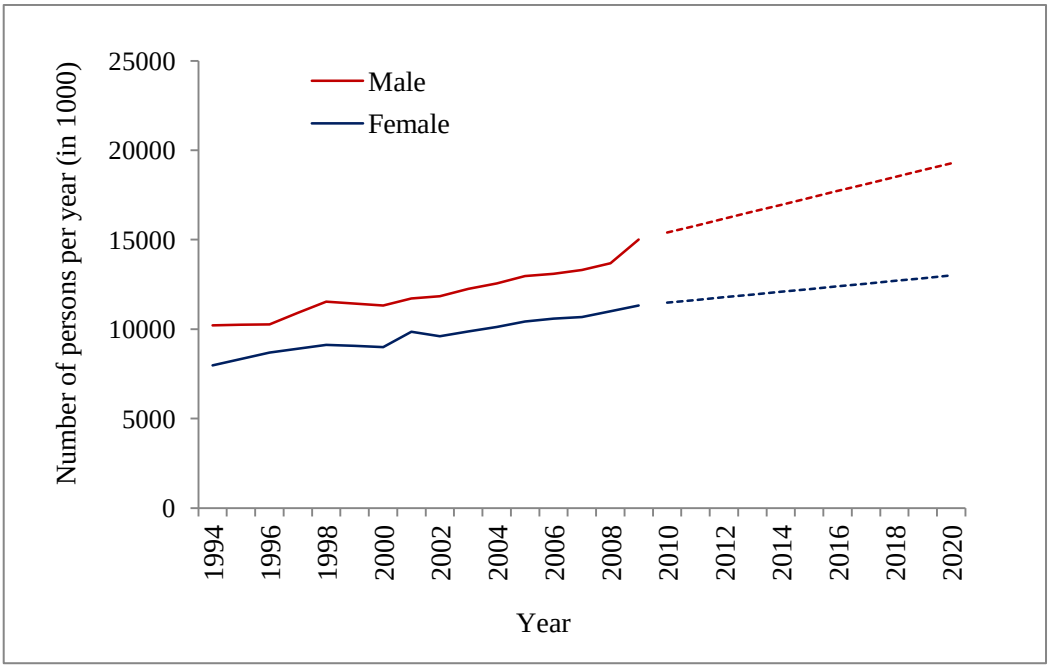


Figure 3.8: Male and female populations in KSA from 1994 to 2020 (thousands).

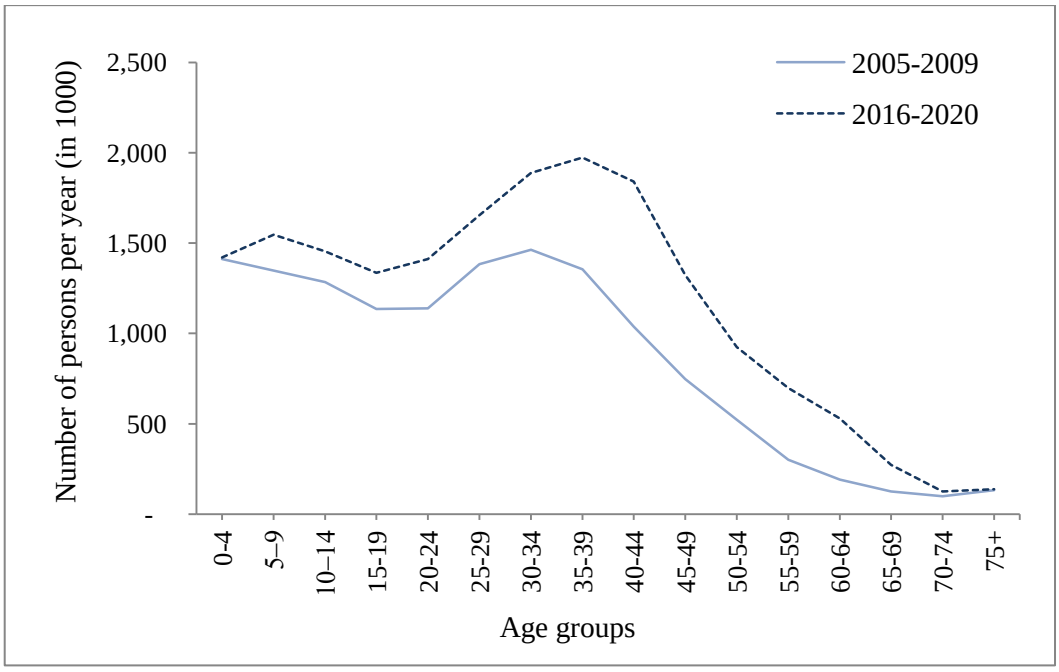


Figure 3.9: Age distribution in thousands of male population in KSA averaged over the period 2005-2009 and the forecast averaged over 2016-2020.

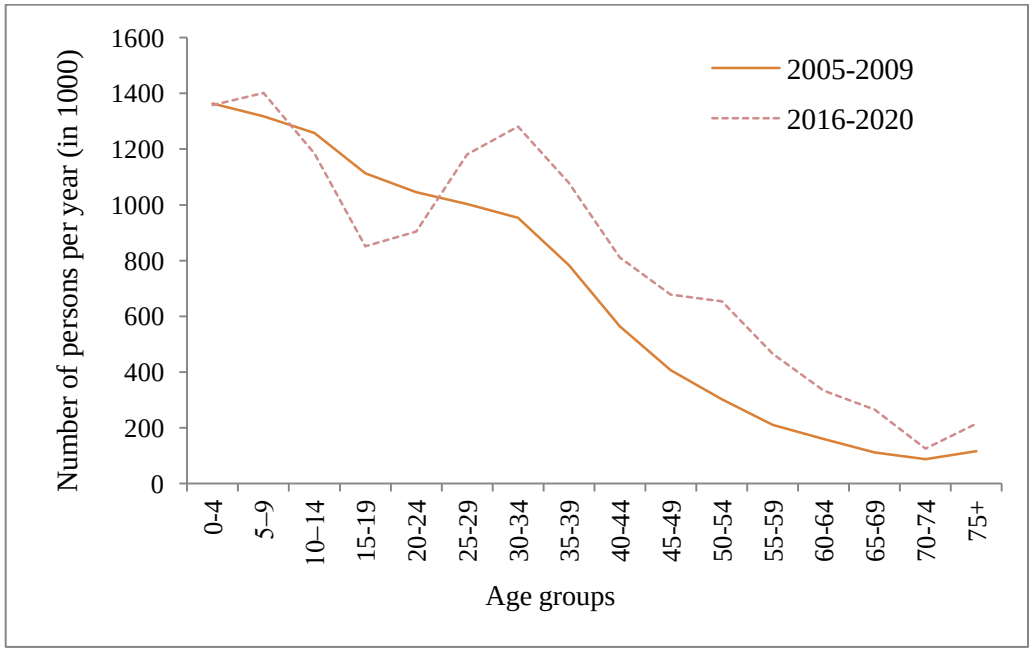


Figure 3.10: Age distribution in thousands of female population in KSA averaged over the period 2005-2009 and the forecast averaged over 2016-2020.

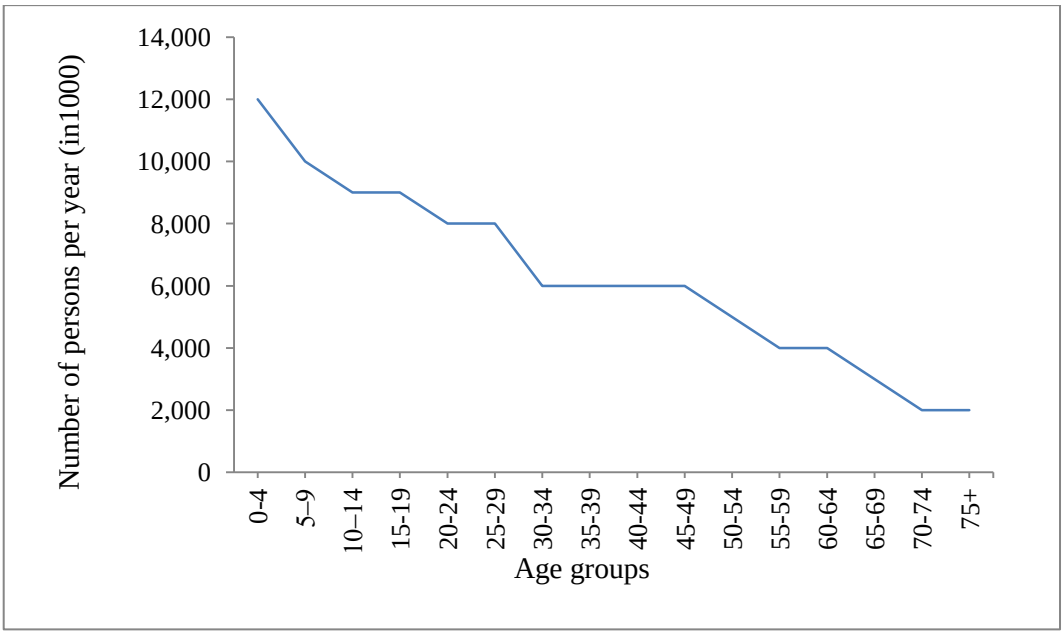


Figure 3.11: Age distribution of the world standard population in 2009.

3.5. Summary

The data collected from 1994 to 2009 for lung cancer from Saudi Cancer Registry (SCR) and Central Department of Statistics (CDS) in the Kingdom of Saudi Arabia (KSA) are presented in four different groups. The groups are Saudi male, Saudi female, non-Saudi male, and non-Saudi female. The lung cancer data collected are the date of diagnosis, gender, ethnicity, type of lung cancer, region, age at diagnosis, date of birth, and the status (dead, alive, or unknown). Population data we have include the number of people at risk from 1994 to 2009 for each age group, gender, and ethnicity. Between 2010 and 2020, the Department of Economic and Social Affairs at the United Nations (2012) made forecasts of population growth for both genders. In addition, data on smoking prevalence by gender, price of imported tobacco, and consumption of tobacco per 1000 tons are presented.

The total number of incident cases reported to the SCR from 1994 to 2009 was 5,966. Overall lung cancer was substantially higher among males than females. Of this 3,487 are Saudi males, 1,145 are non-Saudi males, 1,028 are Saudi females, and 306 are non-Saudi females. In the case of mortality, the total number reported from 1994 to 2009 was 1,755 deaths for males and 486 for females.

These data will be used in the analyses that we carry out in Chapters 4, 5, 6 and 7. For Box-Jenkins methodology in Chapter 4 we use monthly data. The monthly data were also used for dynamic regression modelling of autoregressive model AR(1), distributed lag

models (DLMs), polynomial distributed lag models (PDLs) and autoregressive polynomial distributed lag models (ARPDLs) in Chapter 5. For the age-period-cohort modelling in Chapter 6 we use yearly incidence data using spline functions. In Chapter 7 we use the annual mortality data for the Bayesian dynamic APC models. The use of monthly data helps to address the fact that the dataset is relatively small. Nonetheless, several cancer research studies have used datasets of roughly the same size when making predictions.

CHAPTER 4

PREDICTION OF LUNG CANCER INCIDENCE IN SAUDI ARABIA USING BOX-JENKINS METHODOLOGY

4.0. Introduction

In time series analysis, the Box–Jenkins methodology applies Autoregressive Moving Average (ARMA) or Autoregressive Integrated Moving Average (ARIMA) models to find the best fit of a time series to past values of this time series, in order to make forecasts. Box-Jenkins represents a powerful methodology that addresses trend and seasonality well, see George et al. (1994). ARIMA models have a strong theoretical foundation and provide an effective technique for approximating any stationary process.

4.1. SARIMA (Seasonal ARIMA) Model

A stationary time series x_t is said to follow an autoregressive moving average model of orders *p* and *q*, denoted by ARMA(*p, q*), if it satisfies the following equation

$$x_t - \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots + \alpha_p x_{t-p} = \omega_t + \beta_1 \omega_{t-1} + \beta_2 \omega_{t-2} + \cdots + \beta_q \omega_{t-q} \qquad 4.1$$

where the *α's* and the *β's* are constants such that the model is both stationary and invertible. ω_t is a white noise process.

Equation 4.1 can be written as

$$\phi_p(L) x_t = \theta_q(L) \omega_t \qquad 4.2$$

where $\phi_p(L) = 1 - \alpha_1 L - \alpha_2 L^2 - \cdots - \alpha_p L^p$, $\theta_q(L) = 1 + \beta_1 L + \beta_2 L^2 + \cdots + \beta_q L^q$ and $L^k(x_t) = x_{t-k}$.

The zeros of $\phi_p(L)$ must lie outside the unit circle for stationarity, and for invertibility the zeros of and $\theta_q(L)$ must also lie outside the unit circle.

In case the series are seasonal, the Box-Jenkins methodology proposes multiplicative seasonal models coupled with long-term differencing, if necessary, to achieve stationarity in the mean.

Let *d* be the minimum order for stationarity. Then the resultant stationary series is denoted by $\nabla^d x_t$ where $\nabla = 1 - L$. If $\nabla^d x_t$ follows an ARMA (*p, q*) model, then the original series x_t is said to follow an *autoregressive integrated moving average model of orders p, d and q,* denoted by ARIMA(*p, d, q*). In general, we will write the model as

$$\phi_p(L)(1-L)^d x_t = \theta_q(L)\omega_t \quad 4.3$$

Box and Jenkins (1976) proposed further that a seasonal series of period *S* could be modelled by

$$\Phi_P(L^S)\phi_p(L)\nabla_s^D\nabla^d x_t = \Theta_Q(L^S)\theta_q(L)\omega_t \quad 4.4$$

where ω_t is the usual Gaussian white noise process. The general model in Equation (4.4) is denoted as SARIMA(*p,d,q*) x $(P,D,Q)_S$ and is called a *multiplicative seasonal autoregressive integrated moving average model* . The ordinary autoregressive and moving average components are represented by polynomials $\phi_p(L)$ and $\theta_q(L)$ of orders *p* and *q* respectively, and the seasonal autoregressive and moving average components by $\Phi_P(L^S)$ and $\Theta_Q(L^S)$ of orders P and Q and ordinary and seasonal difference operators by $\nabla^d = (1-L)^d$ and $\nabla_S^D = (1-L^S)^D$. For monthly time series $S = 12$ and for quarterly time series $S = 4$. For estimation of parameters, iterative least squares method is used.

4.2. Model Estimation

In order to fit the model in Equation (4.4), its orders *p, d, q, P, D, Q* and *s* must be determined. One can determine the seasonality period *s* from the nature of the time-series plot and the correlogram. The correlogram of an *s*-period seasonal series exhibits fluctuating movements of the same periodicity as the series.

At each stage of the differencing process, the series is tested for stationarity until it is attained. Here, the Augmented Dickey Fuller (ADF) test shall be used to test for stationarity after each stage of differencing. The AR orders *p* and *P* are estimated as the non-seasonal and the seasonal cut-off lags of the autocorrelation function (ACF) respectively. Similarly the MA orders *q* and *Q* are estimated as the non-seasonal and the seasonal cut-off lags of the partial autocorrelation function (PACF) respectively.

The parameters are thereafter estimated by the use of non-linear optimization techniques because of the involvement of white noise process items in the model. After model fitting the fitted model is usually subjected to residual analysis for validation. All analysis in this work was done using the statistical package R.

In this analysis, we aim to fit a time series SARIMA model to the lung cancer incidence in the Kingdom of Saudi Arabia (KSA). The best-fit model will also be used for forecasting future incidence of lung cancer. The data set contains monthly cases of lung cancer recorded by the Cancer Registry of Saudi Arabia between 1994 and 2009. We are particularly interested in the short-term future forecast of the lung cancer incident cases.

The following analysis mainly focuses on the application of Box-Jenkins SARIMA modelling techniques to estimate the appropriate model that can be used for forecasting future incidence of lung cancer in KSA.

4.3. Analysis

The process consists several stages in an analysis of this type. First, as with any data analysis, we should construct a time plot of the data, and inspect the graph for any anomalies. If, for example, the variability in the data grows with time, it will be necessary to transform the data to stabilize the variance. In such cases, the Box-Cox class of power transformations could be employed. Also, the particular application might suggest an appropriate transformation. The same time plot gives first answers to questions of stationarity or whether the time series show a seasonal pattern.

This is then followed by an identification of the initial model. This, we achieve by establishing seasonality in the dependent series (seasonally differencing it if necessary), and using function plots of the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the dependent time series to decide which (if any) autoregressive or moving average component should be used in the model. Secondly, we estimate the parameters for a tentative model that has been selected. Thirdly, the estimated model is tested or checked for adequacy to determine if it is the best-fit model for the data (this stage includes both residual diagnostics and over-fitting of the initial model). If the estimation is inadequate, we have to return to step one. Lastly, the final best-fit model is then chosen and used to predict future values of the time series.

4.4. Modelling Seasonal Time Series

4.4.1. SARIMA Model Building

Following Section 4.1 above, the data set was plotted to give an initial guess about the data generation process. Figure 4.1 illustrates a time series plot of the original monthly lung cancer incidence data for KSA from January 1994 to December 2009.

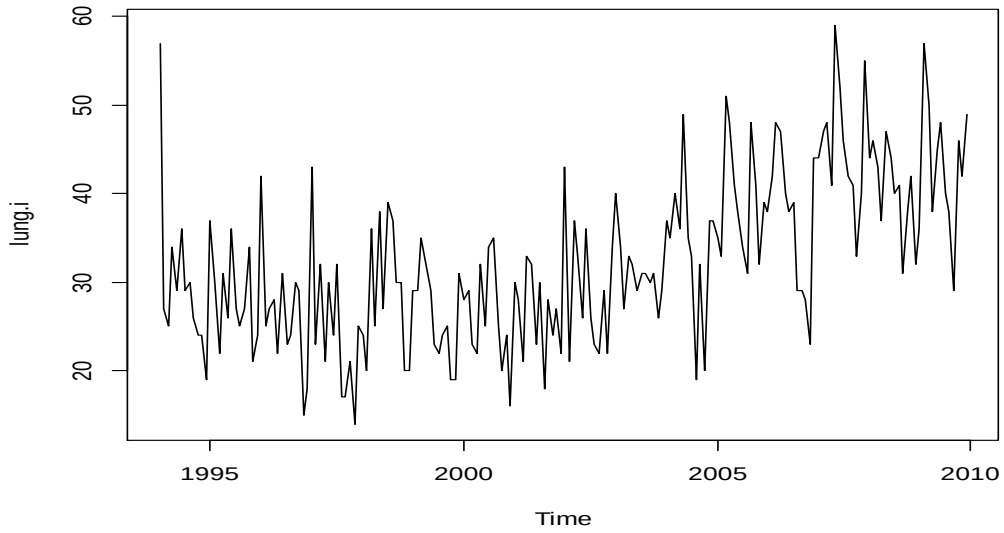


Figure 4.1: Time series plot of the original monthly incidence data.

From the time plot, we could easily observe that many times, the incidence data of lung cancer rises and then drops suddenly within the same year throughout the series from January 1994 to December 2009. One could really see that there is a seasonal pattern where it seems to oscillate with spikes and valleys. Throughout, it seems to have some sort of trend for part of the time in the levels and then increase. Therefore the plot also notifies the presence of profile to it. In addition, the ACF and PACF plots shown in Figure 4.2 of the original series confirm that the dataset is not stationary. Therefore differencing will be necessary so as to attain stationarity.

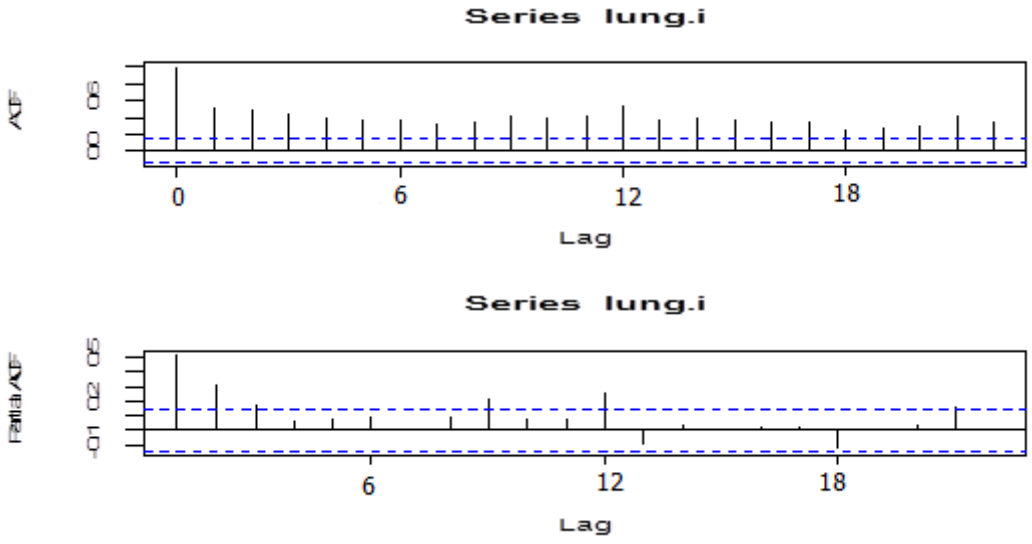
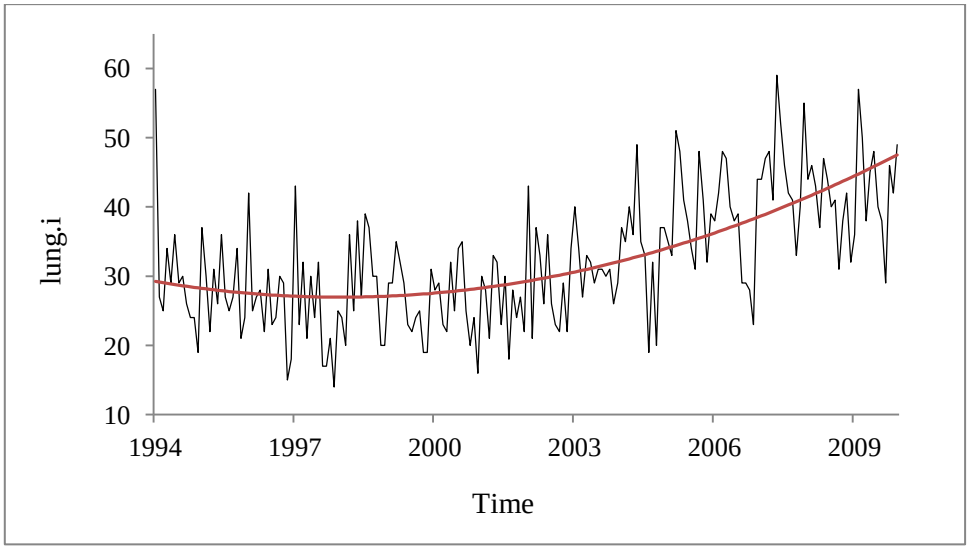


Figure 4.2: ACF and PACF plots of the monthly incidence data.

From the time plot, the series appears to show the presence of a trend. If the series somehow indicate a trend, then there is a possibility of trend-stationarity. A quadratic trend is therefore fit to the data and the de-trended data is plotted in Figure 4.3. Again, this plot did not clearly show that the series is stationary. Since the de-trended data does not clearly indicate stationary, we conclude that the original time series is not trend-stationary. We therefore prefer differencing the time series in order to remove its nonseasonal and seasonal unit roots (Box and Jenkins, 1976).

(a)



(b)

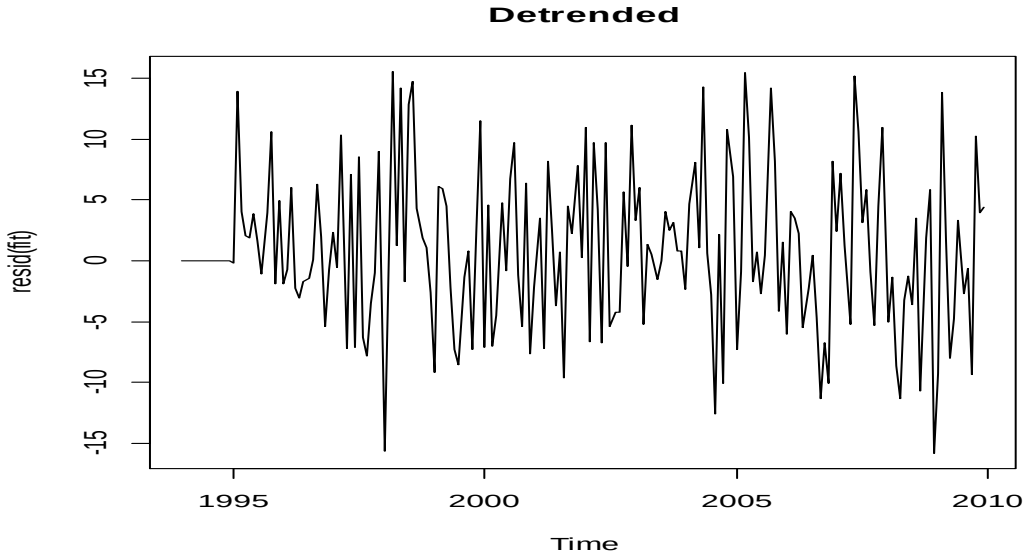


Figure 4.3: (a) Quadratic trend, (b) De-trended data.

4.4.2. Test for Stationarity

Next we consider difference-stationarity. Figure 4.4 shows the time series, autocorrelation function (ACF) and partial autocorrelation function (PACF) plots for the first-difference of the time series.

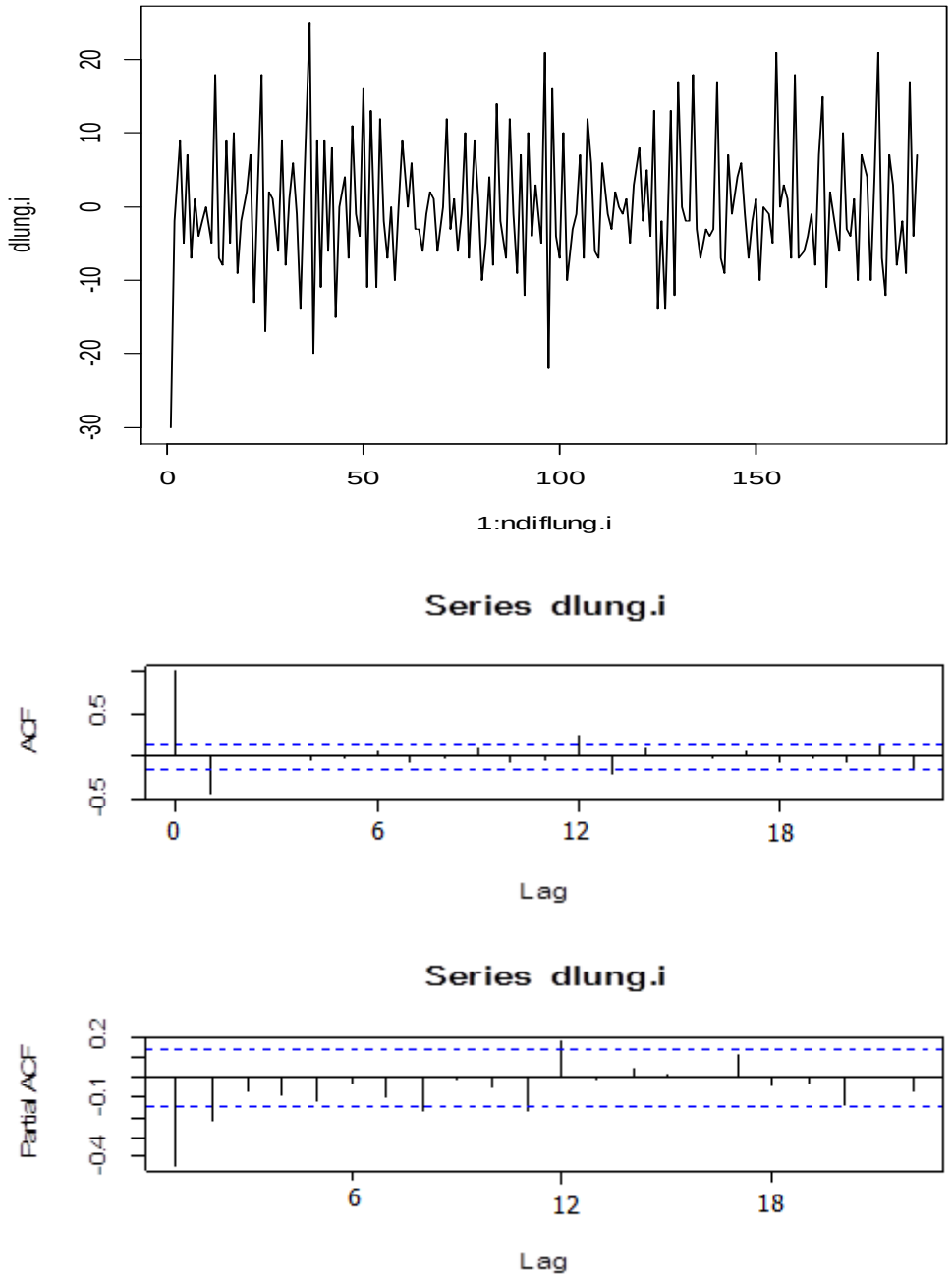


Figure 4.4: First difference of the monthly incidence data - time series, ACF and PACF plots.

An Augmented Dickey-Fuller (ADF) test is used to check for the presence of a unit root in the time series. From the ADF test on the first-difference series, the p-value (0.01) is smaller than 0.05 and therefore we need to accept the alternative hypothesis which is stationary which means again that there is no unit root present in the data (where the null hypothesis assumes that the data is non-stationary). It is therefore clear from the time series plot that this first-differenced time series is stationary (constant mean and approximately constant variance).

The model identification techniques used in Box-Jenkins SARIMA modelling is now applied to the first differenced time series.

4.4.3. Model Identification

Before we can estimate the SARIMA model, we first identify the dependence orders of the model with the aid of autocorrelation function (ACF) and partial autocorrelation functions (PACF).

The output of the differenced data is shown in Figure 4.4. We used a seasonal difference equation $(1 - L^{12})x_t = x_t - x_{t-12}$ and $(1 - L)x_t = x_t - x_{t-1}$ for non-seasonal differencing. The plot shows a transformation of the lung data using the first differencing method to remove the seasonality component in the original series. The pattern move irregularly about its mean value of zero with the variability being approximately stable.

The plot shows clear monthly effect and no obvious trend, so the ACF and PACF of the 12^{th} difference (seasonal differencing) are examined in Figure 4.4. Examining the ACF and the PACF of the difference data, these plots suggest seasonality order of the AR and MA components 2 and 1 respectively.

Therefore we fit

$$\text{SARIMA}(2,1,1)\text{x}(2,1,1)_{12} \qquad 4.5$$

Using the lag operator this model can be written as follows:

$$(1 - \Phi_1 L^{12} - \Phi_2 L^{24})(1 - \phi_1 L - \phi_2 L^2)\nabla_{12}\nabla x_t = (1 + \theta_1 L + \Theta_1 L^{12} + \theta_1\Theta_1 L^{13})\omega_t$$

Using the same procedure, the following other SARIMA models are suggested for comparisons:

- $\text{SARIMA}(2,1,1)\text{x}(1,1,0)_{12}$
- $\text{SARIMA}(2,1,1)\text{x}(1,1,1)_{12}$
- $\text{SARIMA}(2,1,1)\text{x}(1,1,2)_{12}$
- $\text{SARIMA}(2,1,1)\text{x}(0,1,1)_{12}$

- SARIMA(2,1,1)x$(0,1,2)_{12}$
- SARIMA(2,1,1)x$(2,1,2)_{12}$

Using the Maximum Likelihood Estimator the model parameters ϕ, θ, Θ, Φ are estimated. The parameter estimates corresponding to the SARIMA models are shown in Table 4.1. A dash in a box indicates the parameter is not applicable to the respective model.

Table 4.1: Estimated model parameters for SARIMA (p, d, q)x(P, D, Q)$_{12}$.

SARIMA MODEL	ϕ_1 AR(1)	ϕ_2 AR(2)	θ_1 MA(1)	Φ_1 SAR(1)	Φ_2 SAR(2)	Θ_1 SMA(1)	Θ_2 SMA(2)	$\hat{\sigma}^2$
$(2,1,1)x(2,1,1)_{12}$	0.04	0.07	-0.92	-0.04	-0.02	-0.70	-	44.69
$(2,1,1)x(1,1,0)_{12}$	0.05	0.09	-0.96	-0.47	-	-	-	52.06
$(2,1,1)x(1,1,1)_{12}$	0.03	0.07	-0.92	-0.02	-	-0.71	-	44.69
$(2,1,1)x(1,1,2)_{12}$	0.05	0.07	-0.9	-0.86	-	0.17	-0.67	44.23
$(2,1,1)x(0,1,1)_{12}$	0.04	0.07	-0.92	-	-	-0.72	-	44.69
$(2,1,1)x(0,1,2)_{12}$	0.03	0.07	-0.92	-	-	-0.74	0.02	44.69
$(2,1,1)x(2,1,2)_{12}$	0.03	0.07	-0.92	-0.91	-0.12	0.19	-0.58	44.00

Using the standardised residual test, the ACF of the residuals, Normal Q-Q plot of standardised residuals and the Ljung-Box statistic all the seven models were found to be significant.

4.4.4. Model Selection

Fitting the seven models suggested by these observations we obtain the values shown in Table 4.2. Thus the final model was selected using penalty function statistics such as Akaike Information Criteria (AIC, AICc) and Bayesian Information Criterion (BIC).

Table 4.2: Values of AIC, AICc and BIC for the SARIMA Models.

MODEL	AIC	AICc	BIC
SARIMA(2,1,1)x$(2,1,1)_{12}$	4.862	4.876	3.964
SARIMA(2,1,1)x$(1,1,0)_{12}$	4.994	5.006	4.062
SARIMA(2,1,1)x$(1,1,1)_{12}$	4.852	4.865	3.937
SARIMA(2,1,1)x$(1,1,2)_{12}$	4.852	4.866	3.954
SARIMA(2,1,1)x$(0,1,1)_{12}$	4.841	4.854	3.909
SARIMA(2,1,1)x$(0,1,2)_{12}$	4.852	4.865	3.937
SARIMA(2,1,1)x$(2,1,2)_{12}$	4.857	4.872	3.976

From Table 4.2 above, SARIMA (2,1,1)x$(0,1,1)_{12}$ is the best model with the minimum values of Akaike's Information Criteria of AIC, AICc and Bayesian Information Criterion (BIC) statistics. The AIC, AICc and the BIC are good for all the models but the SARIMA(2,1,1)x$(0,1,1)_{12}$ model provided the minimum values and was therefore selected. Therefore we present the estimated model parameters for the best-fit model in Table 4.3 below.

Table 4.3: Estimated parameters of preferred model.

SARIMA(2,1,1)x(0,1,1)	Estimate	Standard Error
AR(1)	0.040	0.085
AR(2)	0.068	0.083
MA(1)	-0.919	0.038
SMA(1)	-0.724	0.070

Hence, SARIMA (2,1,1)x$(0,1,1)_{12}$ is the preferred model, and the fitted model in this case is

$$(1 - \phi_1 L - \phi_2 L^2)\nabla_{12}\nabla\widehat{x_t} = (1 + \theta_1 L + \Theta_1 L^{12} + \theta_1\Theta_1 L^{13})\widehat{\omega_t}$$

$$\left(1 - 0.04_{(0.09)}L - 0.068_{(0.08)}L^2\right)\nabla_{12}\nabla\widehat{x_t}$$

$$= \left(1 - 0.919_{(0.04)}L - 0.724_{(0.07)}L^{12} + (0.919_{(0.04)} * 0.724_{(0.07)})L^{13}\right)\widehat{\omega_t}$$

with $\widehat{\sigma_\omega^2} = 44.69$.

4.4.5. Model Diagnostics

Residual diagnostic for the best-fit model are displayed in Figure 4.5. We note the few outliers in the series as exhibited in the plot of the standardized residuals and their normal Q-Q plot, and a small amount of autocorrelation that still remains (although not at the seasonal lags) but otherwise, the model fits well. Finally, forecasts based on the fitted model for the next 24 months are shown in Figure 4.6.

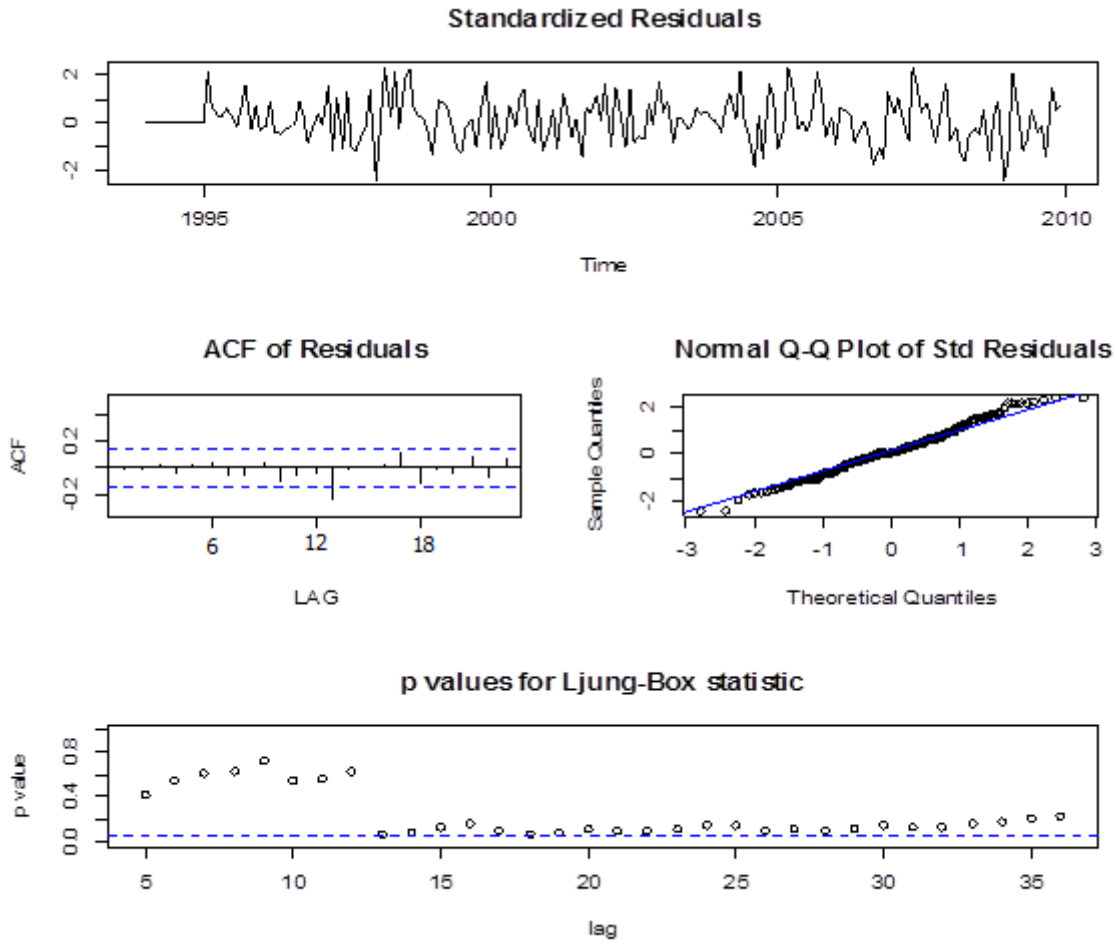


Figure 4.5: Diagnostics for the SARIMA (2,1,1)x(0,1,1)$_{12}$ fit on the lung cancer incidence.

4.5. Forecasting with the SARIMA (2,1,1)x(0,1,1)$_{12}$ model

Forecasts of future incidence of lung cancer are of particular importance to the Ministry of Health of any country.

Using SARIMA (2,1,1)x(0,1,1)$_{12}$, a forecast pattern for the next 24 months ahead of the original data for the period from January, 1994 to December, 2009 was generated.

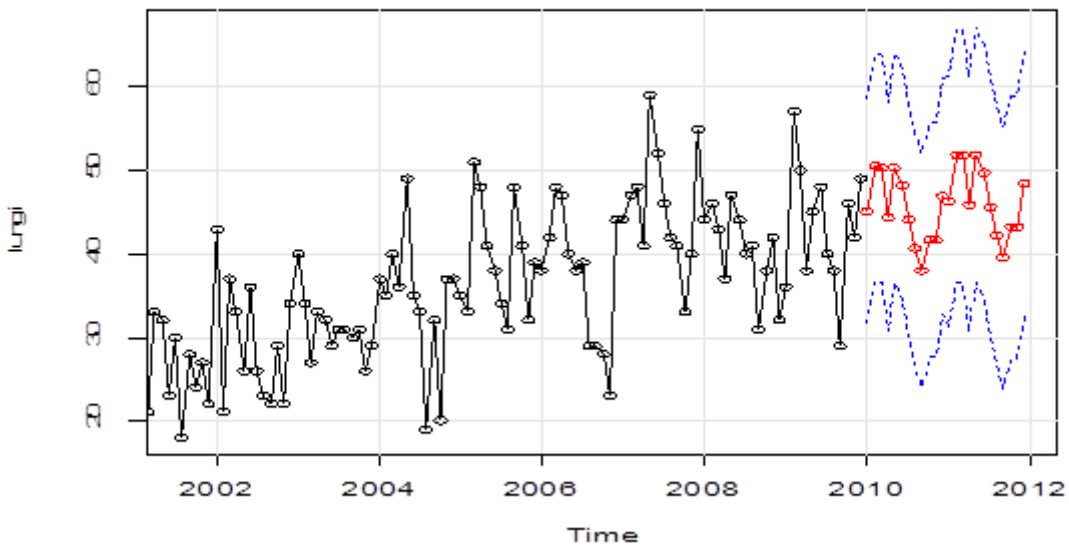


Figure 4.6: Graph of forecast of SARIMA(2,1,1)x$(0,1,1)_{12}$ model.

We may now use the final form of the best-fit SARIMA(2,1,1)x$(0,1,1)_{12}$ model for the time series to estimate future incidence levels. The forecast incidence for the next two years is displayed in Table 4.4 together with the standard errors of the parameter estimates.

Table 4.4: Forecast incidence levels using SARIMA(2,1,1)x$(0,1,1)_{12}$ model.

Month (2010)	Estimate	Standard Error	Month (2011)	Estimate	Standard Error
1	45	6.68	1	46	7.45
2	50	6.73	2	52	7.49
3	50	6.80	3	51	7.54
4	44	6.83	4	45	7.59
5	50	6.86	5	52	7.63
6	48	6.89	6	49	7.66
7	44	6.92	7	45	7.70
8	40	6.95	8	42	7.74
9	38	6.97	9	39	7.78
10	41	7.00	10	43	7.82
11	41	7.03	11	43	7.86
12	47	7.05	12	48	7.90

It is clear from these forecasts that the monthly incidence levels is expected to follow the positive trend visible in the time series plot of the original data.

4.6. Summary

The main aim of this analysis was to determine an appropriate SARIMA model for the lung cancer incidence data in KSA. Particularly, we were interested in forecasting future lung cancer values using this model.

The results of this study indicate that SARIMA model allows for more complex description of the seasonality and autocorrelation structure of the time series and is found to be suitable in predicting the lung cancer incidence in KSA. Based on the minimum AIC, AICc and BIC statistics the best fitted SARIMA model is the SARIMA(2,1,1)x(0,1,1)$_{12}$ expressed as

$$\left(1 - 0.04_{(0.09)}L - 0.068_{(0.08)}L^2\right)\nabla_{12}\nabla\widehat{x_t} = \left(1 - 0.919_{(0.04)}L - 0.724_{(0.07)}L^{12} + (0.919_{(0.04)} * 0.724_{(0.07)})L^{13}\right)\widehat{\omega_t}$$

The model fitted well and provided sensible forecasts for up to 24 months ahead. This is against the backdrop that SARIMA models have shorter period of predicting power (Abraham et al., 2009: Aidoo, 2010).

CHAPTER 5

DYNAMIC REGRESSION MODELLING OF LUNG CANCER INCIDENCE IN SAUDI ARABIA

5.1. Introduction

In this chapter, we model and forecast lung cancer incidence using dynamic regression models with finite and infinite lags. These models involve autoregressive models (ARs), distributed lag models (DLMs), polynomial distributed lag models (PDLs) and autoregressive polynomial distributed lag models (ARPDLs). We give an overview of AR(1) models and how to detect and correct autocorrelation in section 5.2. We outline the implementation and forecasting issues with DLM and PDL models in section 5.5 and section 5.12 respectively and went further to present one-step ahead forecast for the various models in section 5.7. Finally, to evaluate the robustness of the results, we explore ARPDL models in section 5.14 and present our summaries of best models and their forecasts in section 5.17.

5.2. Autoregressive Models

5.2.1. Linear Model of First-order Autoregressive AR(1)

First, we focus on the concept of autocorrelation. In simple terms, autocorrelation means that current values depend on past values. A common starting point of analysis is a simple model of positive first-order AR(1) autocorrelated error-process associated with a regression equation that can be represented by the following two equations:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \qquad 5.1$$

and

$$X_t = S_t \times P_t$$

where Y_t is the incidence at time t (number of cases in month t), X_t is the smoking population in 10,000, S_t is the smoking prevalence and P_t is the population size,

and

$$\varepsilon_t = \rho\, \varepsilon_{t-1} + \nu_t \qquad 5.2$$

where $-1 < \rho < 1$ (or $|\rho| < 1$ in order to avoid unstable behaviour) and $\nu_t \sim N(0, \sigma^2)$. With the model in equation (5.1) above, the errors ε_t are not independently distributed because the previous errors ε_{t-1} influence the current errors whenever ρ is not zero. We say that the errors are autocorrelated. It is important to note that there are three important

consequences when ordinary least squares (OLS) is used to estimate β_0 and β_1 in the presence of an autocorrelated data generation process:

1. The estimates of β_0 and β_1 remain unbiased.
2. The OLS estimated standard errors (SEs) of the estimated coefficients are inconsistent and inference is flawed. Also, they are not asymptotically unbiased.
3. Ordinary least squares does not give the best linear unbiased estimator. A generalized least squares (GLS) procedure is the best (minimum SE) linear unbiased estimator.

Because (OLS) reported SEs are misleading and the OLS estimators are inconsistent if autocorrelation exists, we need to investigate if the errors are autocorrelated (Barretto and Howland, 2006). The following diagnostic procedures are employed:

i. Examining the scatter diagram of OLS residuals plotted against lagged residuals.
ii. Applying the estimated ρ test in which estimated ρ is the slope of the regression of residuals on lagged residuals.
iii. Using the Durbin-Watson (DW) test. Unlike the estimated ρ test, the DW test is often used because it does not suffer from small sample bias.

5.2.2. Detecting Autocorrelation

One simple way of detecting autocorrelation is the autocorrelation function or ACF. Autocorrelation computes and plots the autocorrelations of a time series. Autocorrelation is the correlation between observations of a time series separated by k time units. Suppose we model the total number of lung cancer cases in Saudi Arabia from 1994 to 2009 against smoking population using Equation (5.1) above. The results from Minitab are shown in Table A1 (see Appendix A). The estimated slope from OLS regression through the origin of residuals on lagged residuals is 0.173. Also from the graphs of Figures 5.1 and 5.2, it is concluded that the residuals are serially correlated with positive autocorrelation. Figure 5.1 illustrates the residual plots against time whereas Figure 5.2 shows that the autocorrelation function of the residuals exceed the significance bounds at different lags. Finally, the Durbin-Watson test from Minitab produced the value 1.55538 confirming that autocorrelation exists in the estimated model because according to Asteriou and Hall (2011), the DW statistic is given by $d = 2(1 - \hat{\rho})$. Because ρ by definition ranges from -1 to 1, the range for d will be from 0 to 4. Therefore, we have three main cases:

i. $\rho = 0; d = 2$: therefore, a value of d close to 2 means that there is no evidence of autocorrelation.

ii. $\rho \simeq 1; d \simeq 0$: a strong positive serial correlation indicates that ρ will be close to +1, and thus d will have very low values (close to zero) for positive autocorrelation.

iii. $\rho \simeq -1; d \simeq 4$: similarly, when ρ is close to -1 then d will be close to 4, meaning a strong negative autocorrelation.

From this analysis, we can see that as a rule of thumb, when the DW test statistic is very close to 2 we do not have serial correlation.

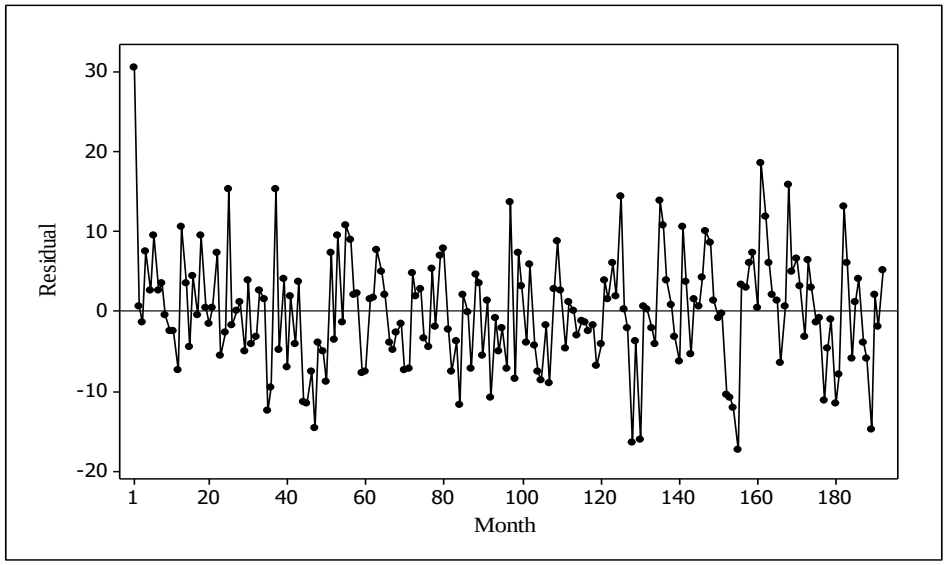


Figure 5.1: Plot of residuals from OLS regression of total cases of lung cancer on smoking population.

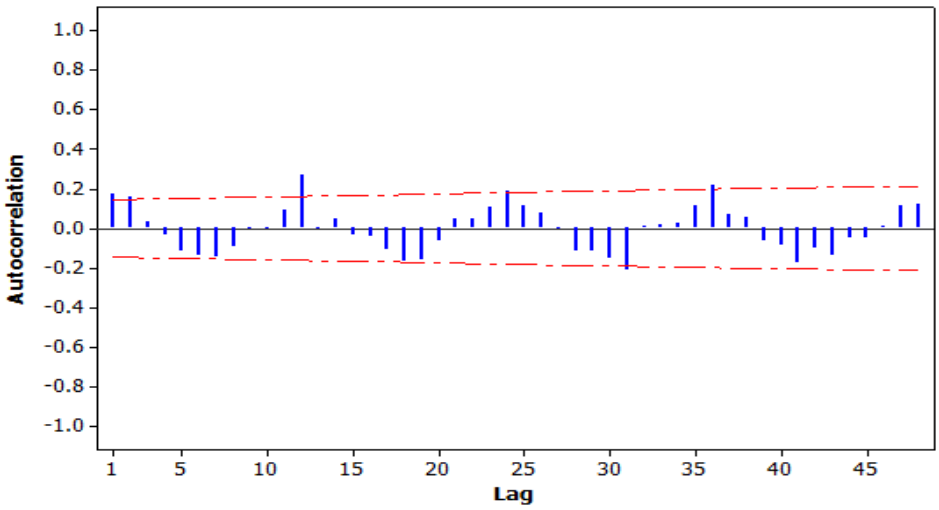


Figure 5.2: Autocorrelation function plot with 95% confidence intervals of the residuals.

5.2.3. Correcting Autocorrelation

Once first-order AR (1) autocorrelated errors are detected, it is possible to correct the model. By appropriately transforming the original data, applying a special formula in section 5.4.4 to the first observation and then running OLS on the transformed data, linear unbiased estimates β_0 and β_1 with minimum SE are found. Applying OLS on appropriately transformed data is called the GLS estimation, and the GLS estimator is the best linear unbiased estimation (BLUE) according to Barretto and Howland (2006).

5.3. Generalized Least Squares

A transformation of Y_t and X_t can be such that the resulting linear model has an independent error structure. We simply transform the model so that we get rid of the ε errors that are systematically related to the previous errors, leaving only the v errors that are independent and normally distributed. This transformation is defined as follows.

By substituting the error-forecasting equation into the equation that generates the observed y:

$$Y_t = \beta_0 + \beta_1 X_t + \rho\, \varepsilon_{t-1} + v_t \qquad 5.3$$

Because each individual Y is generated the same way, lagging equation (5.1) by one period gives

$$Y_{t-1} = \beta_0 + \beta_1 X_{t-1} + \varepsilon_{t-1} \qquad 5.4$$

Multiplying equation (5.4) by the autocorrelation coefficient ρ gives

$$\rho\, Y_{t-1} = \rho\, \beta_0 + \rho\, \beta_1 X_{t-1} + \rho\, \varepsilon_{t-1} \qquad 5.5$$

Subtracting equation (5.5) from equation (5.3) we get that

$$Y_t - \rho\, Y_{t-1} = \beta_0 - \rho\, \beta_0 + \beta_1 X_t - \rho\, \beta_1 X_{t-1} + \rho\, \varepsilon_{t-1} + v_t - \rho\, \varepsilon_{t-1}$$

Rearranging this equation we obtain a model in which the error term is a pure, independently and identically distributed error, v_t:

$$Y_t - \rho\, Y_{t-1} = \beta_0(1-\rho) + \beta_1 X_t - \rho\, \beta_1 X_{t-1} + v_t \qquad 5.6$$

If we define new variables $Y_t^* = Y_t - \rho\, Y_{t-1}$ and $X_t^* = X_t - \rho X_{t-1}$, then we have

$$Y_t^* = \beta_0(1-\rho) + \beta_1 X_t^* + v_t \qquad 5.7$$

Equation (5.7) is known as the transformed model with a well-behaved error term. Since v_t, the disturbance, obeys all the classical conditions by assumption, OLS may be applied to this equation.

However ρ is unknown, so it must be estimated from the regression of residuals on lagged residuals. We then use it to transform the original data to obtain the new

transformed Y_t^* and X_t^*. Note that β_0 and β_1 are the original parameter values of the model. Running OLS on the transformed model is called generalized least squares (GLS). It generates the right SEs as it is the best linear unbiased estimator. Notice as well that when ρ=0, the transformation reduces to the familiar OLS model. Our new model is called generalized least squares because the transformation applied here is one of many possible transformations, which includes OLS as a particular case.

Note that if the regression equation is misspecified or the error process does not follow the AR (1) model, the transformation presented above will not work.

5.4. Iterative Procedures to Estimate ρ

Although the method of generalized differencing seems to be easy to apply, in practice the value of ρ is not known. Therefore, procedures need to be developed to provide us with estimates of ρ and then of the regression model in equation (5.7). Several procedures have been developed, but the most popular ones are Cochrane-Orcutt iterative procedure, Prais-Winsten, and Hildreth-Lu search procedures.

5.4.1. The Cochrane-Orcutt Iterative Procedure

Cochrane-Orcutt (1949) developed an iterative procedure that is described as follows:

1. Estimate the model by OLS and obtain the residuals ε_t .
2. Estimate the first-order autocorrelation coefficient ρ by OLS from
$$\hat{\varepsilon}_t = \rho\, \hat{\varepsilon}_{t-1} + \upsilon_t$$
3. Transform the original data as $Y_t^* = Y_t - \rho\, Y_{t-1}$ and $X_t^* = X_t - \rho\, X_{t-1}$ for t= 2,…, 192. Note that this means that we lose the first observation.
4. Now regress Y_t^* on X_t^* . The constant in this regression will be $\beta_0(1-\rho)$. Generate new residuals from this regression.
5. Regress the new residuals on the lagged new residuals to estimate a new ρ.
6. Go to step 3 and repeat until convergence.

The iterative procedure can be stopped when the estimates of ρ from two successive iterations differ by no more than a preselected (very small) value such as 0.001. The final estimated rho is used to get the estimates of Equation (5.7).

To apply generalized differencing estimation to the total cases of lung cancer in Saudi Arabia from 1994 to 2009 against smoking population, we first need to find an estimate of ρ. We obtain ρ from running a regression of the residuals against lagged residuals obtained from Equation (5.1). We get the results shown in Table A2 (see Appendix A) from which

the ρ coefficient is 0.1726 using STATA 13 statistical package. Then we go to step 3 and repeat the procedure until we get convergence as in Table 5.1 below.

Table 5.1: Cochrane-Orcutt iterative procedure for the best estimated ρ.

Iteration	ρ
1	0
2	0.1726
3	0.1726
4	0.1726

Now, we can see from Table 5.1 that the best estimated ρ when using the Cochrane-Orcutt iterative procedure is 0.1726.

Now to estimate the coefficients, we know that β_1 =0.120, $\beta_0(1-\rho)$ and $\hat{\rho}$= 0.1726. Then $\hat{\beta}_0(1-\hat{\rho}) = -1.344$, so $\hat{\beta}_0 = -1.623$. The final model is

$$Y_t = -1.623 + 0.120X_t \quad \text{and} \quad \varepsilon_t = 0.172\,\varepsilon_{t-1} + \nu_t \text{ where } \nu_t \sim N(0, \sigma^2) \text{ iid.}$$

5.4.2. Prais-Winsten Procedure

Prais-Winsten (1954) is essentially the same as the Cochrane-Orcutt iterative procedure except that we keep the first observation and it does not iterate. Therefore, in order to transform the variables in the first observation we need to apply the following formula to the first observation as follows:

$$Y_1^* = \sqrt{(1-\rho^2)}\; Y_1 \qquad \text{and} \qquad X_1^* = \sqrt{(1-\rho^2)}\, X_1$$

whereas to transform the variables for observations 2 to 192 we use the transformations

$$Y_t^* = Y_t - \rho\, Y_{t-1} \quad \text{and} \quad X_t^* = X_t - \rho\, X_{t-1}.$$

Thence

$$Y_t^* = \beta_0(1-\rho) + \beta_1 X_t^* + \upsilon_t.$$

Suppose that we use the same data but we apply a special formula to the first observation as mentioned earlier. This produced the following iterations for rho (see Table 5.2).

Table 5.2: Prais-Winsten iterative procedure for the best estimated ρ.

Iteration	ρ
1	0.0000
2	0.1726
3	0.1728
4	0.1728

From Table 5.2, the best estimated ρ when using Prais-Winsten iterative procedure is 0.1728. The full results of this are shown in Table A3 (see Appendix A).

Now, we know that $\hat{\beta}_1$= 0.115, $\beta_0(1-\rho)$ and $\hat{\rho}$= 0.173. Then, $\hat{\beta}_0(1-\hat{\rho}) = 0.140$. So $\hat{\beta}_0 = 0.1692$. The final model is

$$Y_t = 0.169 + 0.116\, X_t \quad \text{and} \quad \varepsilon_t = 0.173\, \varepsilon_{t-1} + \nu_t \qquad \text{where } \nu_t \sim N(0, \sigma^2) \text{ iid.}$$

5.4.3. The Hildreth-Lu Search Procedure

Hildreth and Lu (1960) developed an alternative method to the Cochrane-Orcutt iterative procedure as shown in the following.

1. Choose a value for ρ (for example ρ = 0.1), and for this value transform the model as in Equation (5.7) and estimate it by OLS.
2. From the estimation in step 1, obtain the residuals $\widehat{\varepsilon_t}$ and the residual sum of squares (RSS for ρ =0.1). Next choose a different value of ρ (for example ρ = 0.2) and repeat steps 1 and 2.
3. By varying ρ from 0 to 1 in some predetermined systematic way, we can get a series of values for RSS (ρ_i). We choose the ρ for which RSS is minimized and Equation (5.7), which was estimated using the chosen ρ as the optimal solution.

Table 5.3: The Hildreth-Lu search procedure for the best estimated ρ.

Iteration	ρ
1	0.0000
2	0.9999
3	0.5000
4	0.2500
5	0.3750
6	0.3125
7	0.2812
8	0.2656
9	0.2578
10	0.2539
11	0.2519

The best estimated ρ with minimum RSS is shown in Table 5.3 (ρ=0.2519) and the full results from the transformed model are shown in Table A4 (see Appendix A).

Now, we know that $\hat{\beta}_1$= 0.120, $\beta_0(1-\rho)$ and $\hat{\rho}$= 0.2510. Then $\hat{\beta}_0(1-\hat{\rho}) = -1.502$. So

$\hat{\beta}_0 = -2.005$. The final model is

$$Y_t = -2.005 + 0.120X_t \quad \text{and} \quad \varepsilon_t = 0.252\,\varepsilon_{t-1} + \nu_t \quad \text{where} \quad \nu_t \sim N(0, \sigma^2)\ iid.$$

5.4.4. Remark

Generally, as we know, OLS implicitly treats the X value for the intercept term as 1 for each observation. Here, the transformed model, however, has changed the intercept term from 1 to (1-ρ). When estimating this model, either we need to interpret the reported intercept coefficient as an estimate of $\beta_0(1-\rho)$, or we can try the computer software to support the usual intercept in favour of the transformed intercept.

There is an additional sticky detail to consider. How do we transform the first observation? No previous observed value of the independent or dependent variables is available, and thus we cannot apply the formula for the transformed model on the first observation. It turns out that the following formula is the correct transformation to apply to the first observation:

$$Y_t^* = \sqrt{(1-\rho^2)}\ Y_1$$

$$X_t^* = \sqrt{(1-\rho^2)}\ X_1$$

$$\text{Intercept}_1^* = \sqrt{(1-\rho^2)}$$

Intuitively, what this transformation accomplishes is to ensure that the error term in the transformed equation for the first observation has the same spread as the other error terms (e.g. the spread of the υ_t). For more details, see Greene (2000), p.543, or Goldberger (1991), pp. 302-303.

5.5. Distributed Lag Models (DLMs)

5.5.1. Introduction

Distributed lag models are useful because they allow a dependent variable to depend on past values of an explanatory variable at various lags. When the population is increasing, this means that the age distribution will increase over time. This together with the increase in tobacco consumption will lead to a serious problem in the future. The effects of smoking do not occur instantaneously but are spread, or distributed over time. Therefore, decision

makers or action planners should take into account the past or lagged values of the policy variables. Algebraically, we can demonstrate this lag effect by saying that a change in a policy variable X_t has an effect on the dependent or response variables Y_t, Y_{t+1}, ... If we turn this around slightly, then we can say that Y_t is affected by the values of X_t, X_{t-1}, or

$$Y_t = f(X_t, X_{t-1}, X_{t-2}, X_{t-3}, X_{t-4}, \ldots, X_{t-k}) \quad 5.8$$

This distributed lag model is finite as the duration of the effects is a finite period of time, namely k periods. This model is said to be dynamic because it describes the reaction over time. The most important issue here is the lag length, how far back in time must we go. In reality, there are two kinds of lags. First, a finite distributed lag that describes the effects only for a certain and fixed period of time. Second, an infinite distributed lag that describes the effects as lasting and forever. In order to convert Equation (5.8) into a distributed lag model we need a functional form with an error term and then make assumptions about the properties of the error term.

5.5.2. Finite Distributed Lag Models

To model the finite distributed lag, the functional form is assumed to be linear, so that the finite lag model, with an additive error term, is

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 X_{t-3} + \cdots + \beta_k X_{t-k} + \varepsilon_t$$

$$= \alpha + \sum_{i=0}^{k} \beta_i X_{t-i} + \varepsilon_t \quad 5.9$$

where we assume that E (ε_t) = 0, Var (ε_t) = σ^2, and Cov (ε_t, ε_s) = 0 for all s≠ t.

In this model the parameter α is the intercept and β_i is the distributed lag weight to reflect the fact that it measures the effect of changes in past values of X on the expected current value of Y, all other things being equal. Equation (5.9) can be estimated by least squares if the error term ε_t has the usual desirable properties. The question here is, how many lags are required in order to have a correctly specified equation? Or, in other words, what is the optimal lag length?

One way to overcome this is to use a relatively large value for k, estimate the model for $k, k-1, k-2, \ldots$ lags and choose the model with the lowest value of AIC (Akaike Information Criterion), SBC (Schwarz Bayesian Criterion) or any other criterion. However, this procedure will create two kinds of problems:

a) Because of close relationships between the independent variables $X_t, X_{t-1}, .., X_{t-k}$, the model will suffer from multicollinearity.
b) A large number of lag indicates a serious loss of degrees of freedom because we can only choose $k+1$ to n observations.

There are many consequences of collinearity. Firstly, the estimates of least squares are imprecise, meaning that a wide interval estimates will be detected. Secondly, high levels of correlation among the regressors imply multicollinearity, which leads to unreliable and inconsistent coefficient estimates with large variances and standard errors. Thus, because the pattern of lag weights will often be used for policy analysis, decision makers should specify the lag length very carefully since this imprecision may lead to serious problems on decision making.

5.5.3. Short and Long-Run Effects

It is interesting to examine the effect of the βs in Equation (5.9). In order to test for short and long-run effects we should include lags of the dependent and independent variables in the regression model. The main reason for including lags is that we believe that the influence of a variable could extend beyond the period being estimated. The use of lags allows us to find the difference between the short and long run multiplier effects.

As in Equation (5.9), the short run multiplier effect is β_0 since it captures the current effect of any change in X on Y at time t, whereas the long run multiplier is the sum $\beta_0 + \beta_1 + \cdots + \beta_k$ which measures the effect of the permanent change in the value of X.

Notice that when introducing lags, this assumes that not just the current value of the X variable is uncorrelated with the residual, but also all past values of X beyond the lags already included in the model $E(\varepsilon_t | x_{t-i}) = 0$ where $i = 0, \ldots, k-s$ (which changes the definition of exogeneity a little and ensures that the lagged values included in the original model comprise all the possible non-zero dynamic effects of X).

If we assume that the residuals are also uncorrelated with all future values of X this is called *strict exogeneity* $E(\varepsilon_t \mid X_{t+k+s} \ldots X_t \ldots X_{t-k-s}) = 0$ and there may be estimation techniques other than OLS that can be used to estimate dynamic causal effects (Hill et al, 2000).

From empirical studies using real data, it has been shown that short-term forecasts are more reliable than long-term forecasts because the forecast relies more on immediate past observations than long-term observations. Short-term objectives, for example, help decision makers in meeting the long-term objectives, making them an important element of

any decision making. Suppose that the Ministry of Health's long-term goal is to reduce the cases of lung cancer by 10% every year. To do so, it creates a plan that involves a series of short-term forecasts. The Ministry then moves from one short-term objective to the next, knowing that each completed objective brings it closer to its overall goal. Additionally, in some situations, the long-term forecast might fail, which is another reason short-term forecast is important.

To find out the relationship between the dependent variables and one current value of the independent variables to get the short and long-run multiplier of lung cancer incidence in KSA, we regress Y_t on X_t as in the following equation:

$$Y_t = \alpha + \beta_0 X_t + \varepsilon_t$$

Since there are no lags in the estimated model the short and long-run multipliers are the same i.e. $\hat{\beta}_0$ = 0.116 (see Table A5 in Appendix A). Therefore, 1% increase in smoking population suggests an approximately immediate and permanent 12% (43) individual cases increase in lung cancer per month. The value of the Durbin-Watson test indicates that there seems to be first order autocorrelation in the data so standard errors are wrongly estimated, but coefficients are unbiased.

In this case, we have to estimate the model again but the problem is how many lags should be included in the next estimated model. Clearly, we have to increase the number of lags sequentially until the lag values start to become insignificant.

$$Y_t = \alpha + \beta_0 X_t + \varepsilon_t$$
$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t$$
$$..$$
$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \cdots + \beta_k X_{t-k} + \varepsilon_t$$

The main problems here as mentioned earlier are two:

1. Perhaps a limited number of observations in the data set which means "degrees of freedom" problems start to set in and the standard errors of OLS estimates get larger as t-k decreases.
2. More lags increase the risk of multicollinearity, which again increases standard errors and reduces the precision of OLS estimates.

To illustrate this, let us lag the model by one period to see the effect on the model. The full result of fitting the following equation is in Table A6 (see Appendix A).

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t$$

When introducing lags on the smoking population into the model, the short and long-run multipliers are not the same. The short-run (impact) multiplier is $\hat{\beta}_0$ = -1.112 and the long-run multiplier equals $\hat{\beta}_0 + \hat{\beta}_1$(-1.112 + 1.241) = 0.1288. Here, the long-run effect is larger than the short-run effect but seems to agree with the estimate from the first regression without lags. We should note that while the introduction of lags is supposed to reduce autocorrelation, in this case we still appear to have autocorrelation in the model (see Table A6 in Appendix A).

In the next step, we will lag the model, for example by **six periods** instead of one period, to see the effects of adding extra variables to the estimated model. The results of fitting this model are shown in Table A7 (see Appendix A).

The R^2 for the estimated relation is 49.3% and the overall *F*-test value is 24.77. Note however there are big changes to the estimated coefficients and standard errors when adding several lags of smoking population variables and these could be due to multicollineartiy. We have found that none of the smoking population variables are significant according to the p value whereas the rest of the coefficients have changed considerably. The statistical model fits the data quite well and the *F*-test of the joint hypotheses that all distributed lag weights $\beta_i = 0$, $i = 0,...,6$, is rejected at the $\alpha = .05$ level of significance. None of the lag weights is statistically significantly different from zero based on individual *t*-tests, reflecting the fact that the estimates' standard errors are large relative to the estimated coefficients. In addition, the estimated lag weight β_6 is larger than the estimated lag weight for lag 5 and the estimated lag weight β_2 is larger than the estimated lag weight for lag 1. This does not agree with our anticipation that the lag effects of smoking should decrease with time and in the most distant periods should be small and approaching zero.

As a result it is very hard to estimate the short and long run multipliers precisely, yet the short-run (impact) multiplier is -0.46 and the long-run multiplier equals $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_4 + \hat{\beta}_5 + \hat{\beta}_6$ = 0.135. Now we can check the multicollineartiy by looking at the correlation coefficients.

Table 5.4: Correlation coefficients of smoking population x_t, x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4}, x_{t-5} and x_{t-6} with p-values.

	x_t	x_{t-1}	x_{t-2}	x_{t-3}	x_{t-4}	x_{t-5}
x_{t-1}	1.000					
	0.000					
x_{t-2}	0.999	1.000				
	0.000	0.000				
x_{t-3}	0.997	0.999	1.000			
	0.000	0.000	0.000			
x_{t-4}	0.995	0.997	0.999	1.000		
	0.000	0.000	0.000	0.000		
x_{t-5}	0.993	0.995	0.997	0.999	1.000	
	0.000	0.000	0.000	0.000	0.000	
x_{t-6}	0.991	0.993	0.995	0.997	0.999	1.000
	0.000	0.000	0.000	0.000	0.000	0.000

From Table 5.4 above, the correlation coefficients for all smoking prevalence variables are almost collinear. Since the pattern of lag weights will often be used for policy analysis, this imprecision may have serious consequences on the decision making. Therefore, an alternative approach is needed to provide methods that can resolve these difficulties. The typical approach is to impose restrictions regarding the structure of the βs and then reduce from $k+1$ to a few number of parameters to be estimated. Imposing a shape on the lag distribution will reduce the effects of collinearity. Let us assume that the lag weights follow a smooth pattern that can be represented by a low degree polynomial. Two of the most popular methods proposed in 1954 and 1965 respectively for doing this are the Koyck (geometric lag) and the Almon (polynomial lag) transformations.

5.5.4. The Koyck Transformation

Koyck (1954) proposed a geometrically declining scheme for the βs. Therefore, rather than estimate the model with a large number of lags we can transform the data into a more parsimonious form by using the Koyck Transformation procedure.

Begin with a model of Y as a function of X and k lags of X:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 X_{t-3} + \cdots + \beta_k X_{t-k} + \varepsilon_t \qquad 5.10$$

Suppose that in the distributed lagged model (DLM) the effect of variable X_t diminishes as the lag gets larger by an amount λ each period. This is reflected in the size of coefficients such that

$$\beta_i = \beta_0 \lambda^i \quad \text{and} \quad 0 < \lambda < 1,$$

where λ is a fraction, so the larger the value of λ the slower the speed of adjustment.

Substituting $\beta_i = \beta_0 \lambda^i$ into the DLM in Equation (5.10), we get

$$Y_t = \alpha + \beta_0 X_t + \beta_0 \lambda X_{t-1} + \beta_0 \lambda^2 X_{t-2} + \beta_0 \lambda^3 X_{t-3} + \cdots + \beta_0 \lambda^k X_{t-k} + \varepsilon_t \quad 5.11$$

$$= \alpha + \beta_0 (X_t + \lambda X_{t-1} + \lambda^2 X_{t-2} + \lambda^3 X_{t-3} + \ldots + \lambda^k X_{t-k}) + \varepsilon_t \quad 5.12$$

If (5.12) is true at time t it is also true at time t-1, so if we lag Equation (5.12) one time period,

$$Y_{t-1} = \alpha + \beta_0 (X_{t-1} + \lambda X_{t-2} + \lambda^2 X_{t-3} + \lambda^3 X_{t-4} + \cdots + \lambda^k X_{t-k+1}) + \varepsilon_{t-1} \quad 5.13$$

Multiplying Equation (5.13) by λ gives

$$\lambda Y_{t-1} = \lambda\alpha + \beta_0 (\lambda X_{t-1} + \lambda^2 X_{t-2} + \lambda^3 X_{t-3} + \lambda^4 X_{t-4} + \cdots + \lambda^{k+1} X_{t-k+1}) + \lambda\varepsilon_{t-1} \quad 5.14$$

Subtracting Equation (5.14) from Equation (5.12), we obtain

$$Y_t - \lambda Y_{t-1} = \{ \alpha + \beta_0 (X_t + \lambda X_{t-1} + \lambda^2 X_{t-2} + \lambda^3 X_{t-3} + \cdots + \lambda^k X_{t-k}) + \varepsilon_t \} - \{ \lambda\alpha + \beta_0 (\lambda X_{t-1} + \lambda^2 X_{t-2} + \lambda^3 X_{t-3} + \lambda^4 X_{t-4} + \cdots + \lambda^{k+1} X_{t-k+1}) + \lambda\varepsilon_{t-1} \}$$

Simplifying (all lags cancel out) gives

$$Y_t - \lambda Y_{t-1} = (1-\lambda)\alpha + \beta_0 X_t + \varepsilon_t - \lambda\varepsilon_{t-1}$$

Hence

$$Y_t = (1-\lambda)\alpha + \beta_0 X_t + \lambda Y_{t-1} + (\varepsilon_t - \lambda\varepsilon_{t-1}) \quad 5.15$$

Using Equation (5.15), regress Y_t on X_t and Y_{t-1} to generate estimates of β_0 and λ. Use these estimates to compute the coefficients at each lag as well as the original intercept α. This transformation is known as the Koyck transformation. As a result, this model has fewer coefficients to estimate which means less chance of multicollinearity.

Applying the Koyck transformation to the total cases of lung cancer against smoking population, the results are shown in Table A8 in Appendix A. From the estimated equation we can find the coefficient parameters as follows:

$\hat{\beta}_0 = 0.099$, $\hat{\lambda} = 0.176$, and $(1-\hat{\lambda})\hat{\alpha} = -1.182$. So $(1-0.176)\hat{\alpha} = -1.182$ and $\hat{\alpha} = -1.434$

Estimated coefficients of the original equation

$$\beta_i = \beta_0 \lambda^i \quad 5.16$$

are

$$\hat{\beta}_0 = 0.099$$

$$\hat{\beta}_1 = \hat{\beta}_0\,\hat{\lambda} = 0.099\,x\,0.176 = 0.017$$

$$\hat{\beta}_2 = \hat{\beta}_0\,\hat{\lambda}^2 = 0.099\;x\,(0.176)^2 = 0.003$$

$$\hat{\beta}_3 = \hat{\beta}_0\,\hat{\lambda}^3 = 0.099\;x\,(0.176)^3 = 0.001$$

$$\hat{\beta}_4 = \hat{\beta}_0\,\hat{\lambda}^4 = 0.099\;x\,(0.176)^4 = 0.0001$$

$$\hat{\beta}_5 = \hat{\beta}_0\,\hat{\lambda}^5 = 0.099\;x\,(0.176)^5 = 0.00002$$

$$\hat{\beta}_6 = \hat{\beta}_0\,\hat{\lambda}^6 = 0.099\;x\,(0.176)^6 = 0.000003$$

$$\hat{\beta}_7 = \hat{\beta}_0\,\hat{\lambda}^7 = 0.099\,x\,(0.176)^7 = 0.000000$$

$$\hat{\beta}_8 = \hat{\beta}_0\,\hat{\lambda}^8 = 0.099\;x\,(0.176)^8 = 0.000000$$

$$\hat{\beta}_9 = \hat{\beta}_0\,\hat{\lambda}^9 = 0.099\;x\,(0.176)^9 = 0.000000$$

$$\hat{\beta}_{10} = \hat{\beta}_0\,\hat{\lambda}^{10} = 0.099\;x\,(0.176)^{10} = 0.000000$$

$$\hat{\beta}_{11} = \hat{\beta}_0\,\hat{\lambda}^{11} = 0.099\;x\,(0.176)^{11} = 0.000000$$

Hence

$$Y_t = -1.434 + 0.099\,X_t + 0.017X_{t-1} + 0.003\,X_{t-2} + 0.001X_{t-3} + 0.0001X_{t-4} + 0.00002X_{t-5} + 0.000003X_{t-6} \qquad 5.17$$

The short run multiplier β_0 estimate is 0.099 and the parameter λ estimate is 0.176, so the long run multiplier is given by $\hat{\beta}_0/(1-\hat{\lambda})$ = 0.099 / (1- 0.176) = 0.120.

If all the β_i in theses equations are positive, a useful way of summarizing the lag structure is to find the mean lag, given by

$$Mean\ lag = \sum(i\,\beta_i)/\sum\beta_i \qquad 5.18$$

Equation (5.18) is a weighted average of the individual lags in Equation (5.16), with weights given by the relative size of the β_s , for equation (5.17), for example

$$Mean\ lag = 0\left(\frac{0.099}{0.120}\right) + 1\left(\frac{0.017}{0.120}\right) + 2\left(\frac{0.003}{0.120}\right) + 3\left(\frac{0.001}{0.120}\right) + 4\left(\frac{0.0001}{0.120}\right) + 5\left(\frac{0.00002}{0.120}\right) + 6\left(\frac{0.000003}{0.120}\right)$$

= 0.21 periods.

Thus on average a change in *X* takes 0.21 periods before it affects *Y*.

The lag patterns for various values of the parameter λ are illustrated in Figure 5.3.

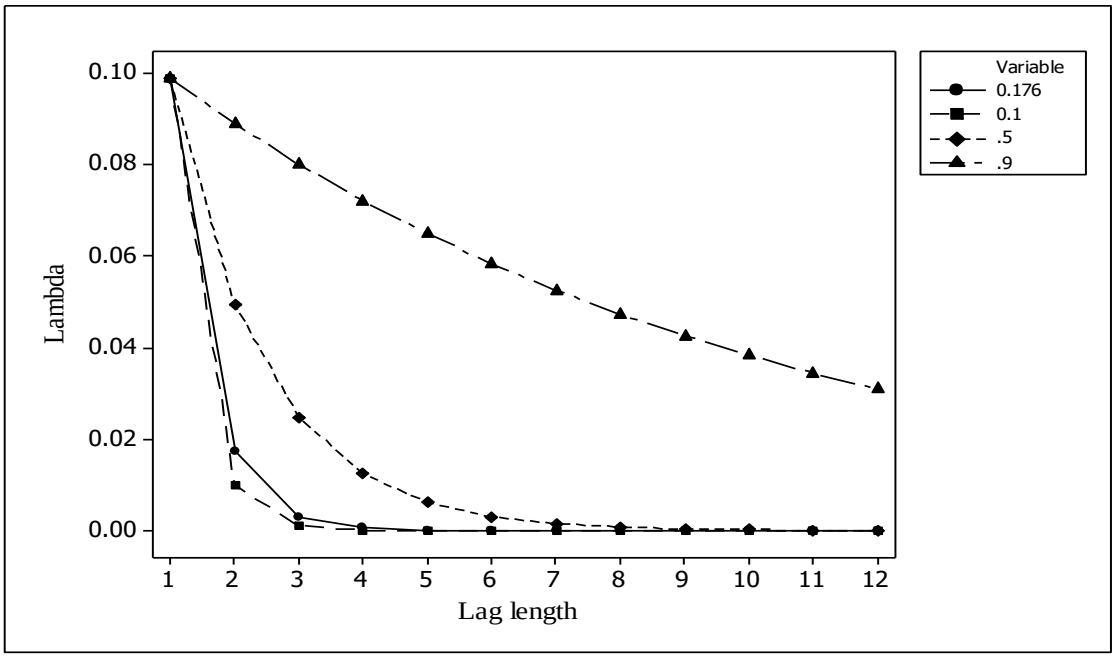


Figure 5.3: Geometric lag coefficients for different values of λ.

5.6. Other Models with Lag Structure

There are several other models for reducing the number of parameters in a distributed lag model. Some of the most important ones are the Pascal lag, the gamma lag, the LaGuerre lag and the Shiller lag. For a full explanation of these models, see Kmenta (1986).

5.7. One-step Ahead Forecasts

The forecasting models included here are four types as follows:

(i) AR (1) model

$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \varepsilon_t$ where $\varepsilon_t \sim N(o, \sigma^2)$,

(ii) Simple linear regression with lagged covariate

$Y_t = \beta_0 + \beta_1 X_{t-1} + \varepsilon_t$ where $\varepsilon_t \sim N(o, \sigma^2)$

(iii) Simple linear regression with lagged covariate and AR(1) errors

$Y_t = \beta_0 + \beta_1 X_{t-1} + \varepsilon_t$,

where $\varepsilon_t = \rho\, \varepsilon_{t-1} + \nu_t$ and $\nu_t \sim N(o, \sigma^2)$.

(iv) Distributed lag model (DLM)

$$Y_t = \alpha + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 X_{t-3} + \cdots + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma^2) \text{ iid.}$$

Now the one-step ahead forecasts are as follows:

5.8. Forecasting with the AR (1) Model

Given data ($Y_1,\ Y_2,\ Y_3,\ Y_4\ , \ldots , Y_T$), the one period ahead optimal forecast is as follows:

$$\begin{aligned}\hat{Y}_{T+1,\ T} &= E(\ Y_{T+1}\ |\Omega_T) \\ &= \alpha_0 + E(\ \alpha_1 Y_T\ |\Omega_T) + E(\ \varepsilon_{T+1}\ |\Omega_T) \\ &= \ \alpha_0 + \alpha_1\, Y_T\end{aligned}$$

In practice, we compute $\hat{Y}_{T+1,\ T} = \ \hat{\alpha}_0 + \ \hat{\alpha}_1 Y_T$ using the estimates.

5.8.1. Forecast Error

The one-step ahead optimal forecast error of AR (1) is

$$Y_{T+1} - \hat{Y}_{T+1,\ T} \ = \ \varepsilon_{T+1}$$

The forecast error variance is

$$Var\ (Y_{T+1} - \hat{Y}_{T+1,\ T}) = \ Var\ (\varepsilon_{T+1}\ ) = \sigma^2$$

5.8.2. Prediction Interval for AR (1) Model

To evaluate the prediction interval we use the normal method:

(i) Assume the forecast error is normally distributed

(ii) Construct the prediction interval (PI) using the following equation

$$\hat{Y}_{T+1} \pm Z_{\alpha/2} \sqrt{\widehat{Var}(\ \hat{Y}_{T+1})}$$

Therefore, the 95% PI is computed as follows:

$$\hat{Y}_{T+1} \pm 1.96\ \sqrt{\widehat{Var}(\hat{Y}_{T+1})}\ .$$

From the results shown in Table A9 in Appendix A, the estimated AR(1) model is

$$Y_t = 14.788 + 0.5462 Y_{t-1} + \varepsilon_t\ .$$

The computed one-step ahead forecast is $\hat{Y}_{T+1} = 41$. The 95% PI for this forecast is (26, 57). This is also calculated as follows:

The AR (1) with T observations has the mean μ = 32.584, α_1 = 0.5462, Y_T= 49, σ^2= 61.2. The AR (1) process is

$$Y_t = \alpha_0 + \alpha_1\ Y_{t-1} + \varepsilon_t\ ,$$

where $\alpha_0 = \mu(1 - \alpha_1)$ so that $\alpha_0 = \mu(1 - \alpha_1)$ = 32.584 (1-0.5462) = 14.788. The one period ahead forecast is 14.788+ 0.5462× 49 = 41 cases of lung cancer. Thus, the one-step ahead forecast is a fixed amount $\alpha_0 + \alpha_1\ Y_{t-1}$ plus the stochastic term ε_t . The fixed amount has a variance of zero, so the variance of the one-step ahead forecast is $\hat{\sigma}^2 = 61.2$. The

plots for one-step ahead forecasts and the residuals are shown in Figure 5.4 and Figure 5.5 respectively.

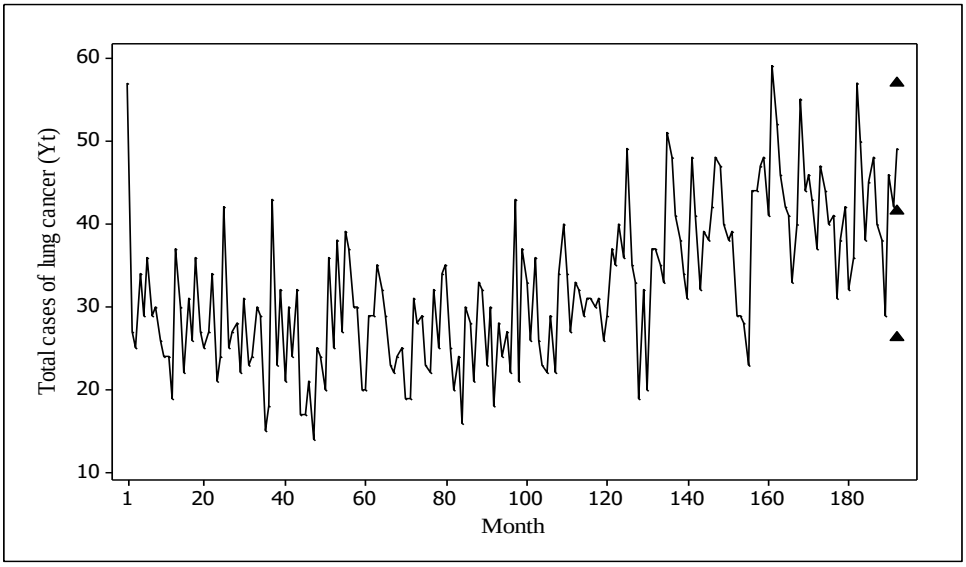


Figure 5.4: One step ahead forecast for AR(1) model with 95% PI.

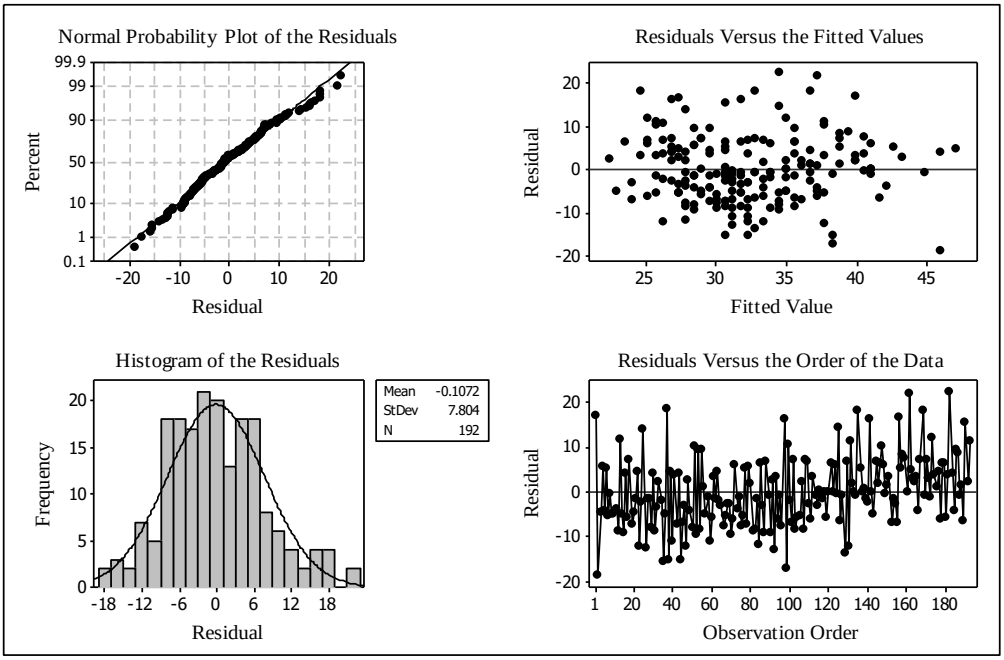


Figure 5.5: Residual plots for AR(1) model for total cases of lung cancer.

5.9. Forecasting with the Linear Regression Model with Lagged Covariate

Having shown how to forecast for the AR(1) model using a one-step ahead forecast, let us forecast for the simple linear regression model with lagged covariate. The model is as follows:

$$Y_t = \beta_0 + \beta_1 X_{t-1} + \varepsilon_t, \ \varepsilon_t \sim N(o, \sigma^2)$$

Given data ($X_1, X_2, X_3, X_4, \ldots, X_T$), the one-step ahead optimal forecast is as follows:

$$\begin{aligned} \hat{Y}_{T+1,\ T} &= E(X_{T+1} \mid \Omega_T) \\ &= \beta_0 + E(\beta_1 X_T \mid \Omega_T) + E(\varepsilon_{T+1} \mid \Omega_T) \\ &= \beta_0 + \beta_1 X_T \end{aligned}$$

In practice, we compute $\hat{Y}_{T+1,\ T} = \hat{\beta}_0 + \hat{\beta}_1 X_T$ using the estimates. The estimated regression equation is as follows:

$$Y_t = \text{-}1.447 + 0.1211\, X_{t-1} + \varepsilon_t .$$

Therefore the one-step ahead forecast when X_T = 377.540 is

$$Y_{T+1} = \beta_0 + \beta_1 X_T$$

$$Y_{T+1} = -1.447 + 0.1211 \times 377.540 = 44 \text{ cases of lung cancer.}$$

5.9.1. Forecast Error

The mean square error of the residuals (variance) is 45.42 as shown in Table A10 in Appendix A.

5.9.2. Prediction Interval

The 95% PI is computed as follows:

$$\hat{Y}_{T+1} \pm 1.96 \sqrt{\widehat{Var}\,(\hat{Y}_{T+1})}$$

$$44 \pm 1.96 \sqrt{45.42}$$

Therefore, the 95% PI is (30, 57). Figure 5.6 and Figure 5.7 show the regression fit and the residual plots respectively.

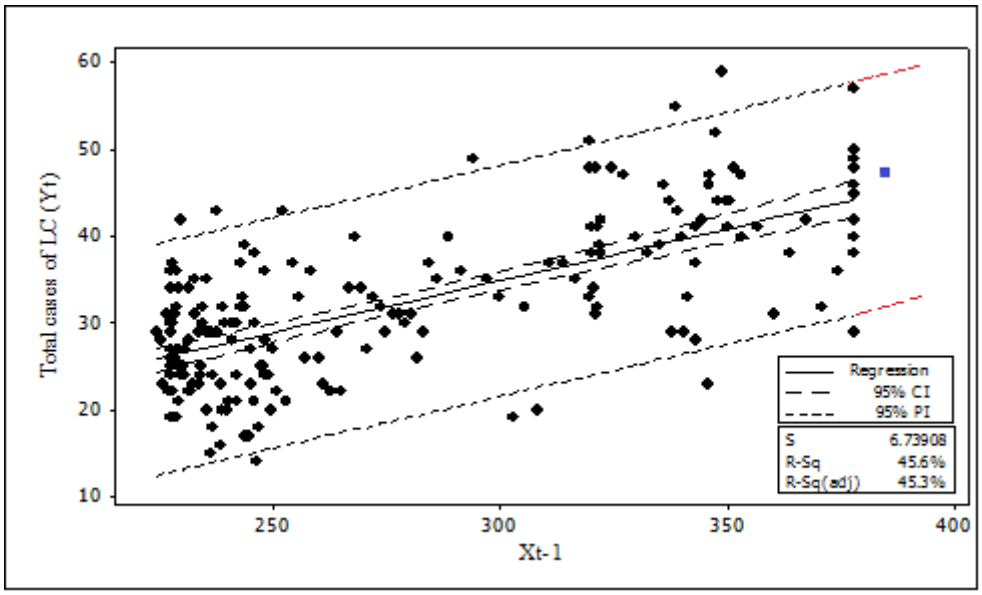


Figure 5.6: Fitted line plot with 95% PI.

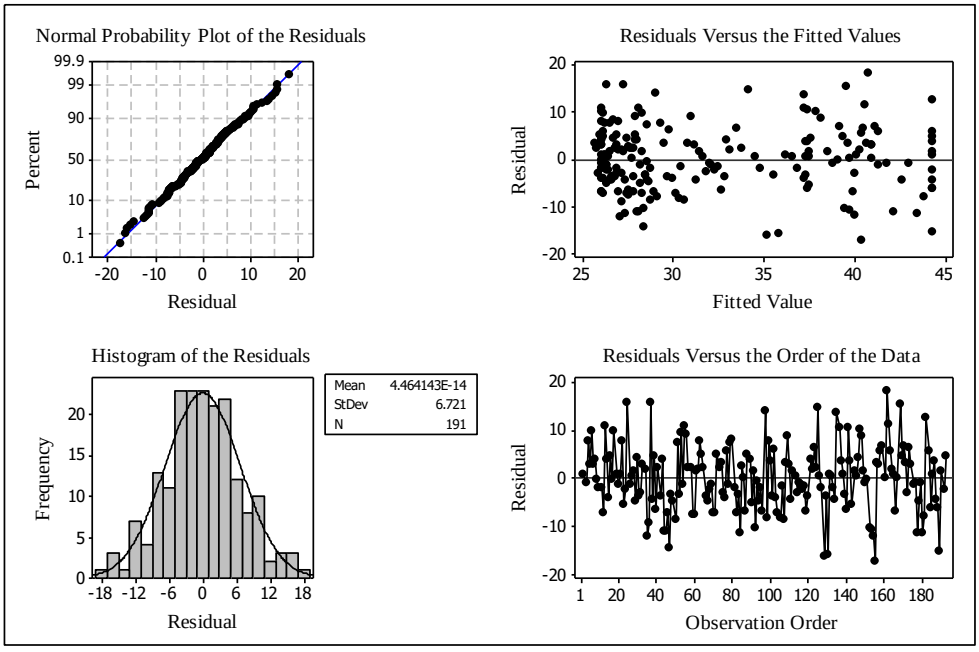


Figure 5.7: Residual plots for linear regression model with lagged covariate.

5.10. Forecasting with the Linear Regression Model with Lagged Covariate and AR (1) Errors

In this section, we consider the case when the errors are correlated and possess first order autocorrelation. The process is as follows:

$$Y_t = \beta_0 + \beta_1 X_{t-1} + \varepsilon_t ,$$

where $\varepsilon_t = \rho\, \varepsilon_{t-1} + \nu_t, \qquad \nu_t \sim N(o, \sigma^2)$.

The estimated AR (1) model when using the Cochrane-Orcutt iterative procedure is as follows:

$$Y_t = \text{-}1.54 + 0.121\, X_{t-1} + \epsilon_t \qquad \text{and} \qquad \epsilon_t = 0.180\epsilon_{t-1} + \nu_t\ .$$

Therefore the one-step ahead forecast when X_T = 377.54 is

$$Y_{T+1} = \beta_0 + \beta_1 X_T$$

$$Y_{T+1} = -1.87 + 0.121 \times 377.54 = 43 \text{ cases of lung cancer.}$$

The mean square error of the residuals (variance) of the one-step ahead forecast is 44.16 as shown in Table A11 in Appendix A.

5.10.1 Prediction Interval

The 95% PI is given as

$$43 \pm 1.96\ \sqrt{44.16}$$

Therefore, the 95% PI is (30, 56). The following Figure 5.8 shows the fitted line plot with 95% PI for the one-step ahead forecast.

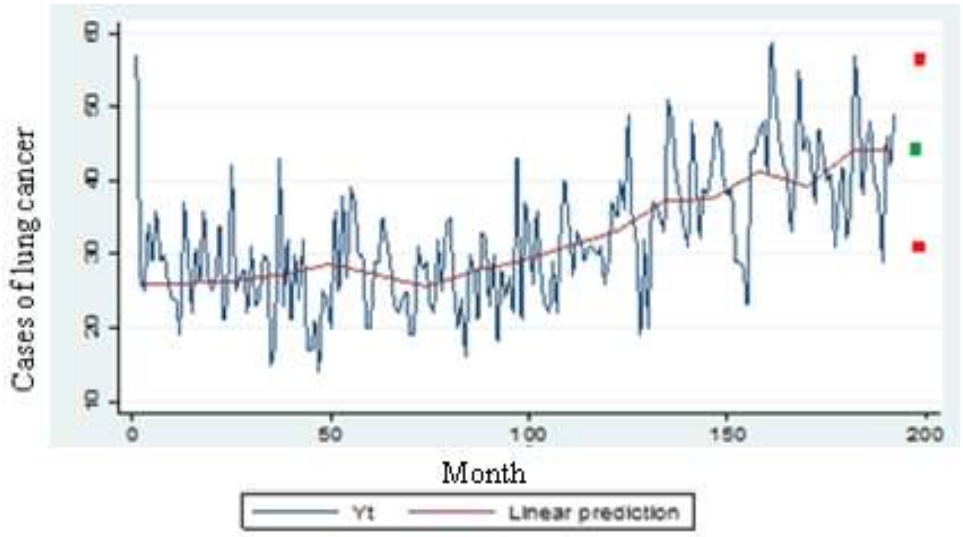


Figure 5.8: Fitted line plot with 95% PI for the one-step ahead forecast.

5.11. Forecasting with the Distributed Lag Model (DLM)

Here, we forecast for the infinite DLM as follows:

$$Y_t = \alpha + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 X_{t-3} + \cdots + \varepsilon_t,$$

$$= \alpha + \sum_{i=1}^{k} \beta_i X_{t-i} + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma^2)\ iid.$$

The estimated DLM using the Koyck transformation is as follows:

$$Y_t = -1.49 + 0.102\, X_{t-1} + 0.169\, Y_{t-1} + \varepsilon_t .$$

Therefore the one-step ahead forecast when X_T = 377.54 and Y_t= 42 is

$$Y_{T+1} = \alpha + \beta_1 X_T + \lambda Y_T$$

$$Y_{T+1} = -1.79 + 0.102 \times 377.54 + 0.169 \times 42 \approx 44 \text{ cases of lung cancer.}$$

5.11.1. Forecast Error

The mean square error of the residuals (variance) is 44.2 as shown in Table A12 in Appendix A.

5.11.2. Prediction Interval

The 95% PI is given as

$$43.80 \pm 1.96\ \sqrt{44.2}$$

Therefore, the 95% PI is (30, 56). Figure 5.9 shows the residual plots as follows:

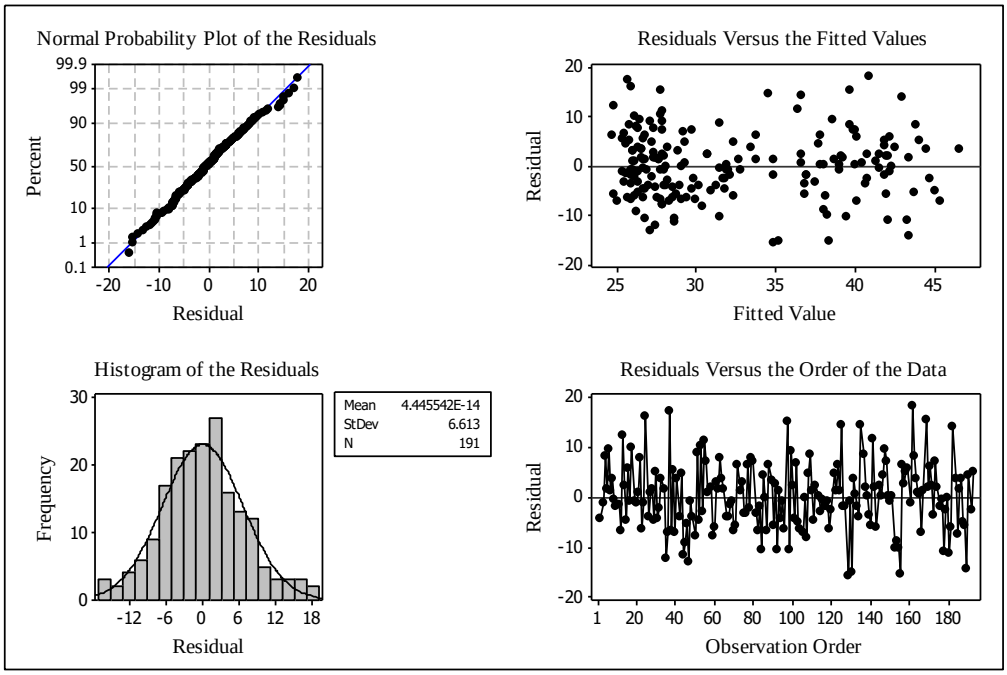


Figure 5.9: Residual plots for DLM model.

To conclude, the above models fail to capture the seasonal pattern in the data so we need to look for more flexible models that can take into account the seasonal effect. Thus, we use polynomial distributed lag models to capture the delays in the time series. We use polynomial distributed lag models because they reduce the amount of data needed to estimate time series phenomena where the numbers of observations available are limited and the number of significant lags are large.

5.12. Polynomial Distributed Lag Models (PDLs)

5.12.1. Introduction

The Koyck transformation may yield to seriously misleading results if one of the explanatory variables in a distributed lag model is not independent of the stochastic disturbance term. Therefore, the OLS estimators may be inconsistent and biased even if the sample size is increased indefinitely. In addition, the Koyck geometric lag is very restrictive in some situations. For instance, if we assumed that the coefficients increase at first and then decline or they follow a cyclical pattern, then the Koyck transformation fails in this case. We therefore need an efficient procedure to correct this problem and hence, the Almon procedure.

To apply the PDL model to the total cases of lung cancer in Saudi Arabia against smoking population data, we have to take into account the following issues. Firstly, the maximum length of the lag k has to be selected. Davidson and Mackinnon (1993) suggested that the best way is to specify the lag length first, by choosing a very large value of k and then seeing whether the fit of the model deteriorates significantly when it is reduced without imposing any restrictions on the shape of the distributed lag. In this case we have to assume that there is a true number of lag lengths and as soon as we underestimate the lag length we will mislead the model to be biased and when we increase the lag length to be more than enough it will increase the risk of multicollinearity. Alternatively, we can use one of the criteria such as Akaike or Schwarz information criterion to choose the appropriate lag length. Secondly, we can specify the order of the polynomial by at least one more than the number of turning points in the curve relating the β_i to i . However, the choice of polynomial degree remains largely subjective if we do not know the number of turning points.

5.12.2. Finite Lags: The Polynomial Lag Model

Almon (1965) developed polynomial lags to approximate inverted U-shaped or even more complicated lag distributions that have a finite rather than an infinite maximum lag. Almon suggested that the immediate impact might be less than the impact after several months, or years. Also after reaching its maximum, the policy effect diminishes for the remainder of the finite lag. In this procedure, we must know how many lags (k) we should include in our model as well as the degree of polynomial (r). Thus, we denote the polynomial distributed lag by PDL(k, r).

Consider the estimation of the equation

$$Y_{jt} = \alpha + \beta_1 X_{jt-1} + \beta_2 X_{jt-2} + \beta_3 X_{jt-3} + \beta_4 X_{jt-4} + \cdots + \beta_k X_{jt-k} + \varepsilon_t \quad 5.19$$

which may be written as

$$Y_{jt} = \alpha + \sum_{i=1}^{k} \beta_k X_{jt-i} + \varepsilon_t$$

and

$$X_{jt-1} = S_{jt-1} \times P_{jt-1}$$

where Y_{jt} is the incidence at time t (number of cases in month t), X_{jt} is the smoking population in 10,000 in month t, S_{jt} is the smoking prevalence and P_{jt} is the population size. Note that when $j = 1$ we refer to males and when $j = 0$ we refer to females.

Almon assumes that the relationship between the β coefficients in Equation (5.19) is approximated by a suitable degree of polynomial r, such as

$$\beta_i = \gamma_0 i^0 + \gamma_1 i^1 + \gamma_2 i^2 + \gamma_3 i^3 + \ldots + \gamma_r i^r \quad 5.20$$

From the lag scheme of our dataset as shown earlier in Figure 3.2, we shall restrict ourselves to third-order polynomial for example. Thus,

$$\beta_i = \gamma_0 i^0 + \gamma_1 i^1 + \gamma_2 i^2 + \gamma_3 i^3 \quad 5.21$$

By substituting Equation (5.21) into Equation (5.19) we obtain

$$\begin{aligned} Y_{jt} &= \alpha + \sum_{i=1}^{k} (\gamma_0 i^0 + \gamma_1 i^1 + \gamma_2 i^2 + \gamma_3 i^3) X_{jt-i} + \varepsilon_t \\ &= \alpha + \gamma_0 \textstyle\sum_{i=1}^{k} X_{jt-i} + \gamma_1 \sum_{i=1}^{k} i\, X_{jt-i} + \\ &\quad \gamma_2 \textstyle\sum_{i=1}^{k} i^2 X_{jt-i} + \gamma_3 \sum_{i=1}^{k} i^3 X_{jt-i} + \varepsilon_t \end{aligned} \quad 5.22$$

Defining

$$Z_{0t} = \sum_{i=1}^{k} X_{jt-i}$$

$$Z_{1t} = \sum_{i=1}^{k} i\, X_{jt-i}$$

$$Z_{2t} = \sum_{i=1}^{k} i^2 X_{jt-i}$$

$$Z_{3t} = \sum_{i=1}^{k} i^3 \, X_{jt-i}$$

and factorizing the γ_is we obtain :

$$Y_{jt} = \alpha + \gamma_0 Z_{0t} + \gamma_1 Z_{1t} + \gamma_2 Z_{2t} + \gamma_3 Z_{3t} + \varepsilon_t \qquad 5.23$$

Now we regress Y_{jt} on the created variables Z_{0t}, Z_{1t}, Z_{2t} and Z_{3t} from the original X_{jt} variables. The polynomial coefficients are then estimated by applying an ordinary least squares (OLS) procedure. As soon as the coefficients are estimated by Equation (5.23), the original β's can then be estimated from Equation (5.21).

What the polynomial approximation has done is to reduce the number of parameters that have to be estimated from $k + 1$ in Equation (5.19) to just five in Equation (5.23). There is a similar reduction in the number of explanatory variables in the estimating equation. The procedure can therefore substantially reduce any multicollinearity problems that might arise in the estimation of Equation (5.19).

In summary, two models are considered: dynamic regression of total cases of lung cancer on total smoking population (Model I) and dynamic regression of total cases of lung cancer on smoking population separately for males and females (Model II) in Section 5.13 and Section 5.15 respectively to see the overall effects of the population classes and budgeting costs for lung cancer in KSA. We use the Almon procedure illustrated by Davidson & MacKinnon (1993) and Maddala & Lahiri (2009, pp 526-533).

5.13. Model I: Dynamic Regression of Total Cases of Lung Cancer on Total Smoking Population

5.13.1: Choosing the Lag Length with OLS

First, let us estimate the unrestricted distributed lag model by running an OLS regression on Equation (5.19) by following the advice of Davidson & MacKinnon (1993) and Maddala & Lahiri (2009, pp 526-533). They both suggested that we need to settle the question of lag length first by starting with a very large value of k and then see whether the fit of the model deteriorates significantly when k reduces without imposing any restriction on the shape of distributed lag. Having specified the best lag length k, we can specify the order of the polynomial r by starting with a very large value of r and then check whether the fit of the model deteriorates significantly when r is reduced.

There are 192 observations and we decided to estimate 36 lagged coefficients. For forecasting purposes, assume that cases (Y_t) depend on smoking population in previous month X_{t-1} and the preceding 35 months as in Equation (5.24) below. The implicit assumption is that the maximum time lag between smoking population X_t and the total cases of lung cancer Y_t is one month. The model is given by

$$Y_t = \alpha + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \cdots + \beta_{36} X_{t-36} + \varepsilon_t \quad 5.24$$

Let us run a regression on the original Equation (5.24) using our monthly data of the total cases of lung cancer Y_t and the total smoking population X_{t-i}.

Using EViews8 software package, the results are presented as in Davidson & MacKinnon (1993) and Maddala & Lahiri (2009, pp 526-533) are shown in Table 5.5. We have run the regression 36 times using different lags, starting from lag 36 to lag 1. Then, we checked where the fit of the models deteriorates significantly.

Table 5.5: Choosing the best lag length from OLS.

coefficient of	Lag							
	31	30	29	28	27	26*	25	24
x_{t-1}	-2.060	-2.070	-2.081	-2.101	-2.054	-2.007	-2.084	-2.065
x_{t-2}	3.294	3.304	3.315	3.335	3.290	3.212	3.291	3.270
x_{t-3}	-0.446	-0.446	-0.446	-0.512	-0.515	-0.484	-0.485	-0.485
x_{t-4}	-1.187	-1.187	-1.058	-0.927	-0.927	-0.927	-0.927	-0.926
x_{t-5}	1.259	1.196	0.937	0.871	0.872	0.872	0.873	0.872
x_{t-6}	-2.324	-2.198	-2.067	-2.067	-2.067	-2.067	-2.068	-2.068
x_{t-7}	0.866	0.803	0.802	0.802	0.801	0.801	0.802	0.802
x_{t-8}	1.608	1.608	1.608	1.608	1.609	1.609	1.610	1.609
x_{t-9}	0.577	0.577	0.578	0.578	0.578	0.578	0.577	0.577
x_{t-10}	-4.696	-4.696	-4.697	-4.697	-4.698	-4.697	-4.697	-4.696
x_{t-11}	7.215	7.221	7.219	7.228	7.214	7.207	7.229	5.729
x_{t-12}	-6.400	-6.405	-6.403	-6.411	-6.397	-6.391	-4.435	-1.369
x_{t-13}	2.512	2.502	2.492	2.473	2.516	2.064	-2.076	-3.626
x_{t-14}	1.289	1.298	1.308	1.327	0.044	0.994	3.160	3.141
x_{t-15}	-1.774	-1.774	-1.773	-1.238	1.320	0.822	0.821	0.821
x_{t-16}	0.400	0.400	0.593	-0.520	-1.841	-1.841	-1.841	-1.840
x_{t-17}	0.231	0.464	0.075	0.652	0.653	0.653	0.654	0.653
x_{t-18}	-1.931	-2.420	-2.224	-2.224	-2.224	-2.224	-2.225	-2.225
x_{t-19}	1.675	1.931	1.931	1.931	1.931	1.931	1.931	1.932
x_{t-20}	1.326	1.327	1.327	1.326	1.327	1.327	1.328	1.327
x_{t-21}	-0.272	-0.272	-0.272	-0.272	-0.272	-0.272	-0.273	-0.273
x_{t-22}	-3.366	-3.366	-3.367	-3.367	-3.367	-3.367	-3.367	-3.366
x_{t-23}	6.581	6.598	6.594	6.621	6.574	6.555	6.621	4.110
x_{t-24}	-10.035	-10.051	-10.047	-10.073	-10.027	-10.008	-6.871	-1.773
x_{t-25}	9.991	9.990	9.989	9.988	9.991	9.184	2.590	
x_{t-26}	-3.040	-3.039	-3.039	-3.038	-5.058	-3.393		
x_{t-27}	-4.277	-4.277	-4.276	-3.302	0.858			
x_{t-28}	4.050	4.050	4.160	2.145				
x_{t-29}	-1.298	-0.819	-1.042					
x_{t-30}	0.882	-0.112						
x_{t-31}	-0.516							
sum of coefficients	0.133	0.135	0.134	0.137	0.132	0.131	0.136	0.131
$\bar{R}^2$	0.456	0.464	0.466	0.473	0.472	0.476	0.469	0.448

There are several features of the lag distribution in the above table. The adjusted R-squared increased gradually until we use a lag of 26. The sum of the coefficients also increases steadily except for lag 27. As a result, it appears that a lag distribution using 26 lags is appropriate. The main problem with the OLS estimates is that, no matter how many lags we include, the Durbin-Watson (DW) test shows positive correlation. This can be seen from the following table of the DW test for different lengths of the lag distribution:

Table 5.6: The Durbin-Watson statistic.

Length of lag	DW
20	1.69
24	1.61
26	1.55
28	1.54
32	1.51
36	1.43

From Table 5.6, this suggests a typical symptom of collinearity and we should be estimating some more general dynamic models, allowing for autocorrelated errors.

Table 5.7: The best-unrestricted least squares (OLS) model with 26 lags.

Variable	Coefficient	p-value
C	-3.97	0.28
x_{t-1}	-2.00	0.13
x_{t-2}	3.21	0.28
x_{t-3}	-0.48	0.88
x_{t-4}	-0.92	0.77
x_{t-5}	0.87	0.78
x_{t-6}	-2.06	0.52
x_{t-7}	0.80	0.80
x_{t-8}	1.60	0.62
x_{t-9}	0.57	0.85
x_{t-10}	-4.69	0.15
x_{t-11}	7.20	0.02
x_{t-12}	-6.39	0.05
x_{t-13}	2.06	0.54
x_{t-14}	0.99	0.77
x_{t-15}	0.82	0.81
x_{t-16}	-1.84	0.60
x_{t-17}	0.65	0.85
x_{t-18}	-2.22	0.53
x_{t-19}	1.93	0.58
x_{t-20}	1.32	0.70
x_{t-21}	-0.27	0.93
x_{t-22}	-3.36	0.34
x_{t-23}	6.55	0.07
x_{t-24}	-10.00	0.01
x_{t-25}	9.18	0.01
x_{t-26}	-3.39	0.05

R-squared	0.558597	Mean dependent var	32.77108
Adjusted R-squared	0.476032	S.D. dependent var	9.408990
S.E. of regression	6.810757	Akaike info criterion	6.822671
Sum squared resid	6447.711	Schwarz criterion	7.328838
Log likelihood	-539.2817	Hannan-Quinn criter.	7.028127
F-statistic	6.765567	Durbin-Watson stat	1.549204
Prob(F-statistic)	0.000000		

After fitting the OLS model with 26 lags, we determine whether all the necessary model assumptions are valid before performing any forecast. If there are any violations, subsequent inferential procedures may be invalid resulting in faulty conclusions. Therefore, it is crucial to perform appropriate model diagnostics.

The fitted model is shown in Figure 5.10 together with residual diagnostic plots. This is followed by the distribution of the series in the histogram with a complement of standard descriptive statistics displayed along with the histogram as shown in Figure 5.11. This figure shows that the skewness of the normal distribution is approximately -0.02. The Jarque-Bera is a test statistic for testing whether the series is normally distributed. The test statistic measures the difference of the skewness and kurtosis of the series with those from the normal distribution. Under the null hypothesis of a normal distribution, the Jarque-Bera statistic is distributed as with 2 degrees of freedom. The reported probability is the probability that a Jarque-Bera statistic exceeds (in absolute value) the observed value under the null hypothesis—a small probability value leads to the rejection of the null hypothesis of a normal distribution. Thus, we do not reject the null hypothesis of the normal distribution with p-value of 0.99 at the 5% level and conclude that the model is normally distributed.

Leverage plots are graphical methods used to diagnose any potential failures of the underlying assumptions of a time series model. We use leverage plots to spot near collinearity between the terms. As we can see from Figure 5.12, the points are compressed towards the vertical line indicating collinearity between the terms. Therefore we look for an adequate model that is more flexible and parsimonious.

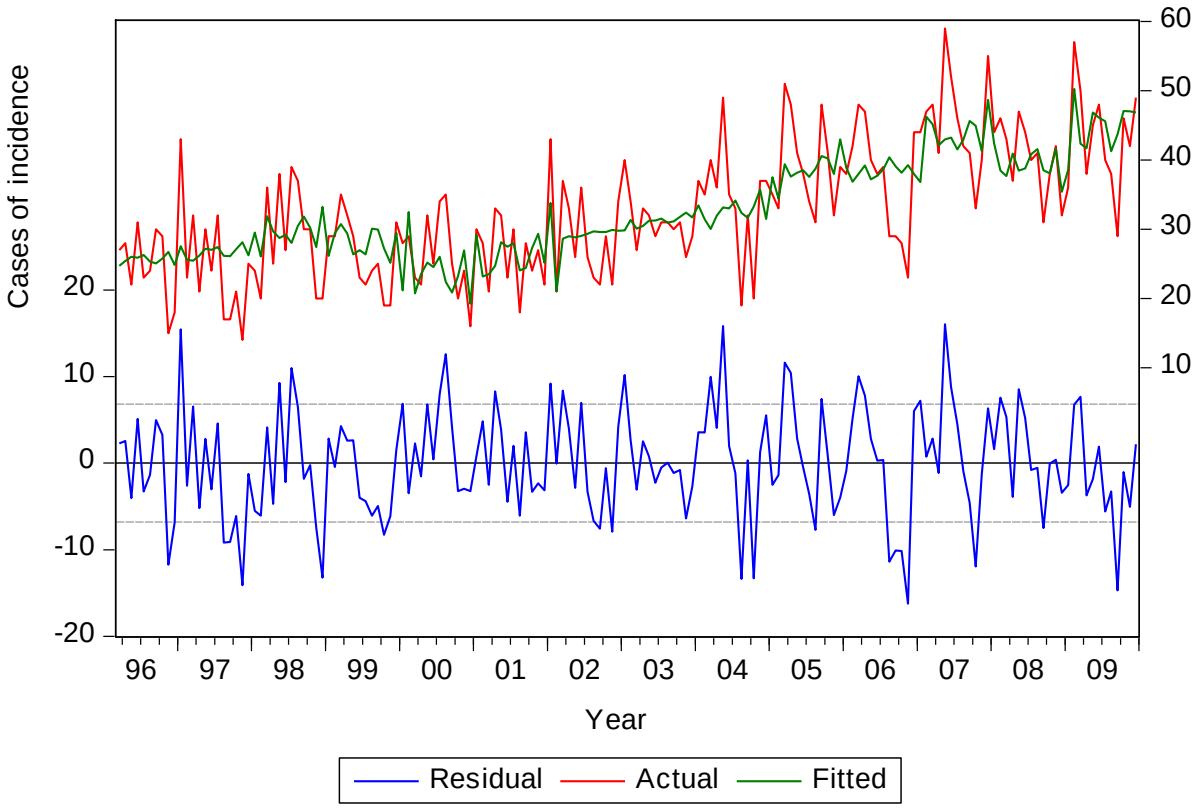


Figure 5.10: Fitted and residual plots for the best OLS model of lung cancer cases per month from 1994 to 2009.

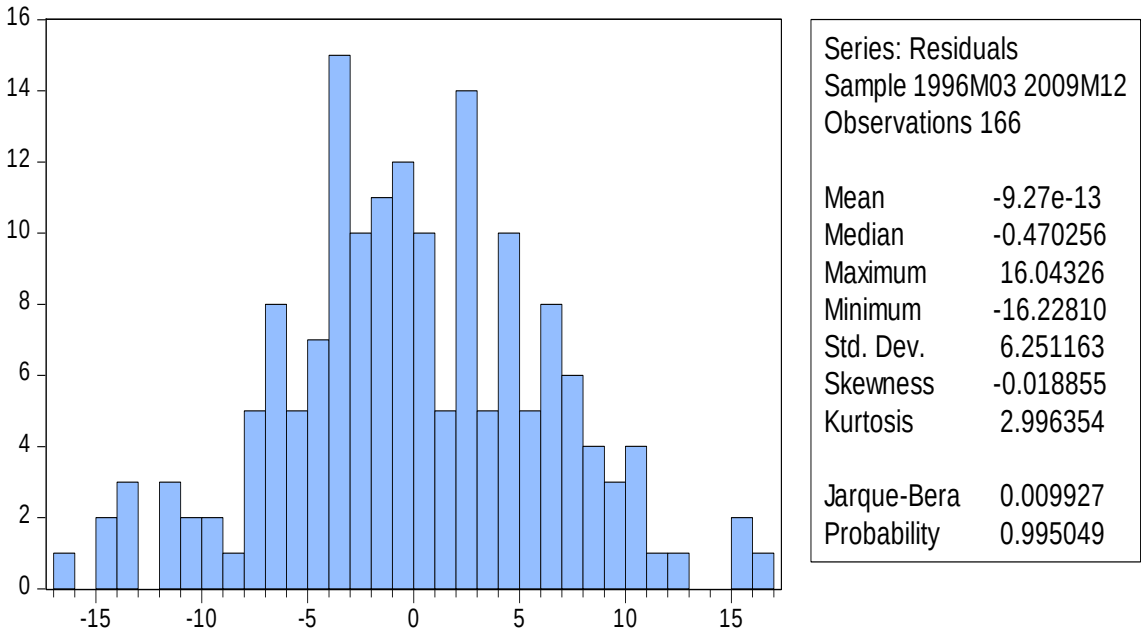


Figure 5.11: Normality plot of the best OLS model of lung cancer cases per month from 1994 to 2009.

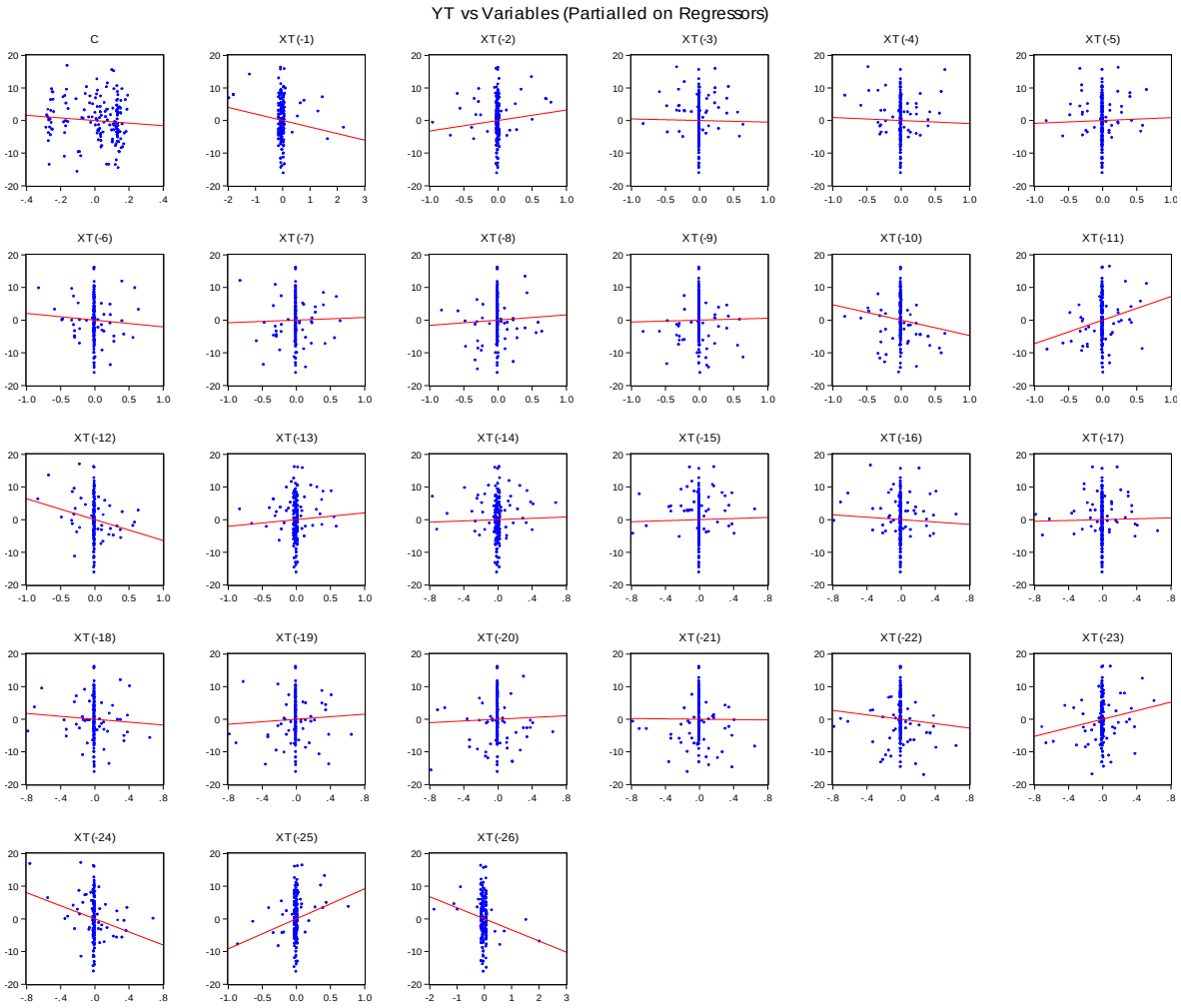


Figure 5.12: Leverage plots for the stability of diagnostics of the best OLS model of lung cancer cases per month from 1994 to 2009.

5.13.2. Choosing the Degree of the Polynomial

As we discussed earlier, the lag length was specified as 26. The next step is to specify the degree of the polynomial by starting with a high-degree polynomial and then we decrease it until we obtain a satisfactory fit, or until one of the hypothesis is rejected (Maddala & Lahiri (2009, pp 526-533)). So we started with a polynomial of degree nine and decreased it until we obtained a satisfactory fit as shown in Table 5.8.

The Durbin Watson (DW) test for first order autocorrelation in regression residuals is the most widely applied tests in time series analysis. A significant test statistic indicates possible mis-specification of the underlying model as well as warning of the invalidity of traditional tests of parameter restrictions. However, the DW test is not inconclusive. Only the boundaries suggested initially by Durbin and Watson were because the precise distribution depends on the observed regressor matrix, which can be address very easily in

most statistical software. In addition, there are generalizations of the DW test to higher lags. So neither inconclusiveness nor limitation of lags is an argument against the Durbin-Watson test (Kleiber and Zeileis, 2008). Therefore, we mainly use the adjusted R-squared values to compare between the models.

Table 5.8: Choosing the degree of the polynomial.

	Equation								
	1			2			3		
coefficient	9th order	t ratios	p-value	8th order*	t ratios	p-value	7th order	t ratios	p-value
Z_{0t}	-0.101424	0.50	0.50	-0.128174	-0.89	0.37	-0.097536	-0.70	0.49
Z_{1t}	-0.146874	0.25	0.25	-0.083218	-1.01	0.31	-0.073998	-0.91	0.37
Z_{2t}	0.018934	0.50	0.50	0.027774	1.14	0.26	0.01194	0.64	0.53
Z_{3t}	0.011895	0.30	0.30	0.00475	1.18	0.24	0.003662	0.94	0.35
Z_{4t}	-0.000539	0.61	0.61	-0.000959	-1.14	0.26	-0.000182	-0.51	0.61
Z_{5t}	-0.000273	0.38	0.38	-6.96E-05	-1.30	0.20	-4.61E-05	-0.95	0.34
Z_{6t}	5.04E-06	0.71	0.71	1.10E-05	1.07	0.29	6.62E-07	0.40	0.69
Z_{7t}	2.33E-06	0.45	0.45	2.90E-07	1.34	0.18	1.65E-07	0.92	0.36
Z_{8t}	-1.46E-08	0.79	0.79	-3.93E-08	-1.01	0.31			
Z_{9t}	-6.59E-09	0.51	0.51						
$\bar{R}^2$	0.481093			0.482945			0.482851		
σ^2	7120.451			7140.808			7187.888		
DW	1.699036			1.701843			1.699121		

First, we test the coefficient of Z_{9t} at the 5% level and we do not reject the hypothesis that it is zero (p=0.51). Next, we test the coefficient of Z_{8t}, also we do not reject the hypothesis that its coefficient is zero (p=0.31). We therefore compare the adjusted R-squared values for the three models and their corresponding DW statistics to select the best order for the polynomial. From Table 5.8, the eighth-order polynomial is appropriate due to its highest adjusted R-squared and DW statistic. Hence, the model PDL(26,8) as illustrated in the following formula

Y_t = -4.25 - 0.128174×i^0 - 0.083218×i^1 + 0.027774× i^2 + 0.00475× i^3- 0.000959× i^4 -
0.0000696×i^5 -0.0000110×i^6 +0.000000290 × i^7 - 0.0000000393×i^8

The results from this model are shown in Table 5.9.

Table 5.9: Results of restricted least squared PDL(26,8) model.

Variable	Coefficient	t-Statistic	p-value
C	-4.25	-1.17	0.24
Z_{0t}	-0.12	-0.89	0.37
Z_{1t}	-0.08	-1.01	0.31
Z_{2t}	0.02	1.13	0.26
Z_{3t}	0.00	1.18	0.24
Z_{4t}	-0.00	-1.13	0.26
Z_{5t}	-0.00	-1.29	0.20
Z_{6t}	0.00	1.06	0.29
Z_{7t}	0.00	1.33	0.18
Z_{8t}	-0.00	-1.01	0.31

R-squared	0.511148	Mean dependent var	32.77108
Adjusted R-squared	0.482945	S.D. dependent var	9.408990
S.E. of regression	6.765679	Akaike info criterion	6.719952
Sum squared resid	7140.808	Schwarz criterion	6.907422
Log likelihood	-547.7561	Hannan-Quinn criter.	6.796047
F-statistic	18.12389	Durbin-Watson stat	1.701843
Prob(F-statistic)	0.000000		

What the polynomial approximation has done is to reduce the number of parameters that have to be estimated from 26 to just 9 in the restricted equation. Therefore, the procedure reduced any multicollinearity problems that might arise in Equation 5.24.

The fitted model is shown in Figure 5.13 together with residual diagnostic plots. This is followed by the distribution of the series in the histogram with a complement of standard descriptive statistics displayed along with the histogram (see Figure 5.14). The p-value (p=0.85) of the Jarque-Bera test is not less than 0.05 for a 5% significance level and hence we do not reject the null hypothesis that the model is normally distributed. Figure 5.15 shows leverage plots of the residuals. We can see that the residuals are not collinear but the fitted model does not reflect the seasonal nature of the data. Therefore, we fit a new autoregressive polynomial distributed lag (ARPDL) model in the next section.

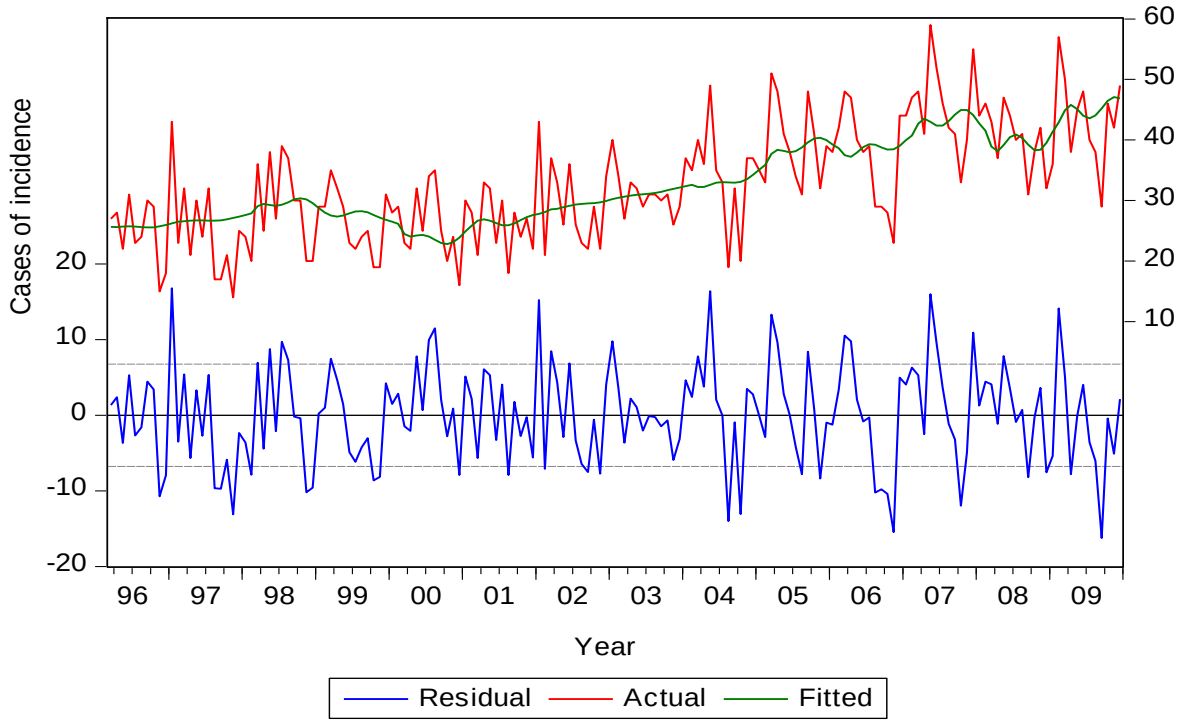


Figure 5.13: Fitted and residual plots for the best PDL(26,8) model of lung cancer cases per month from 1994 to 2009.

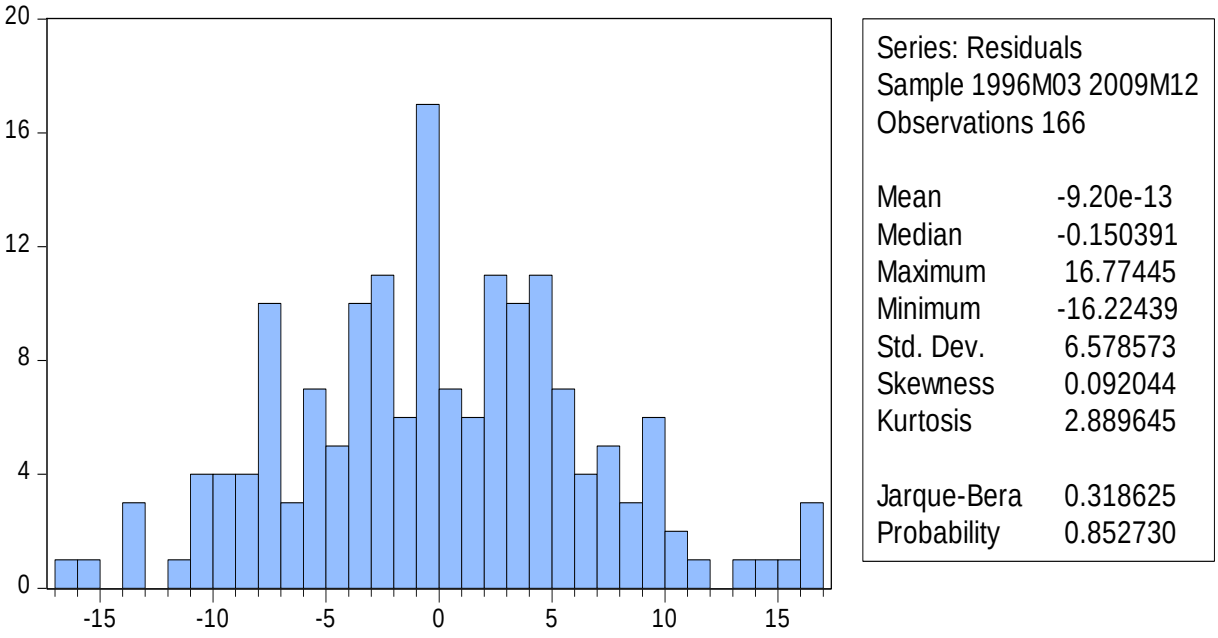


Figure 5.14: Normality plot of the best PDL(26,8) model of lung cancer cases per month from 1994 to 2009.

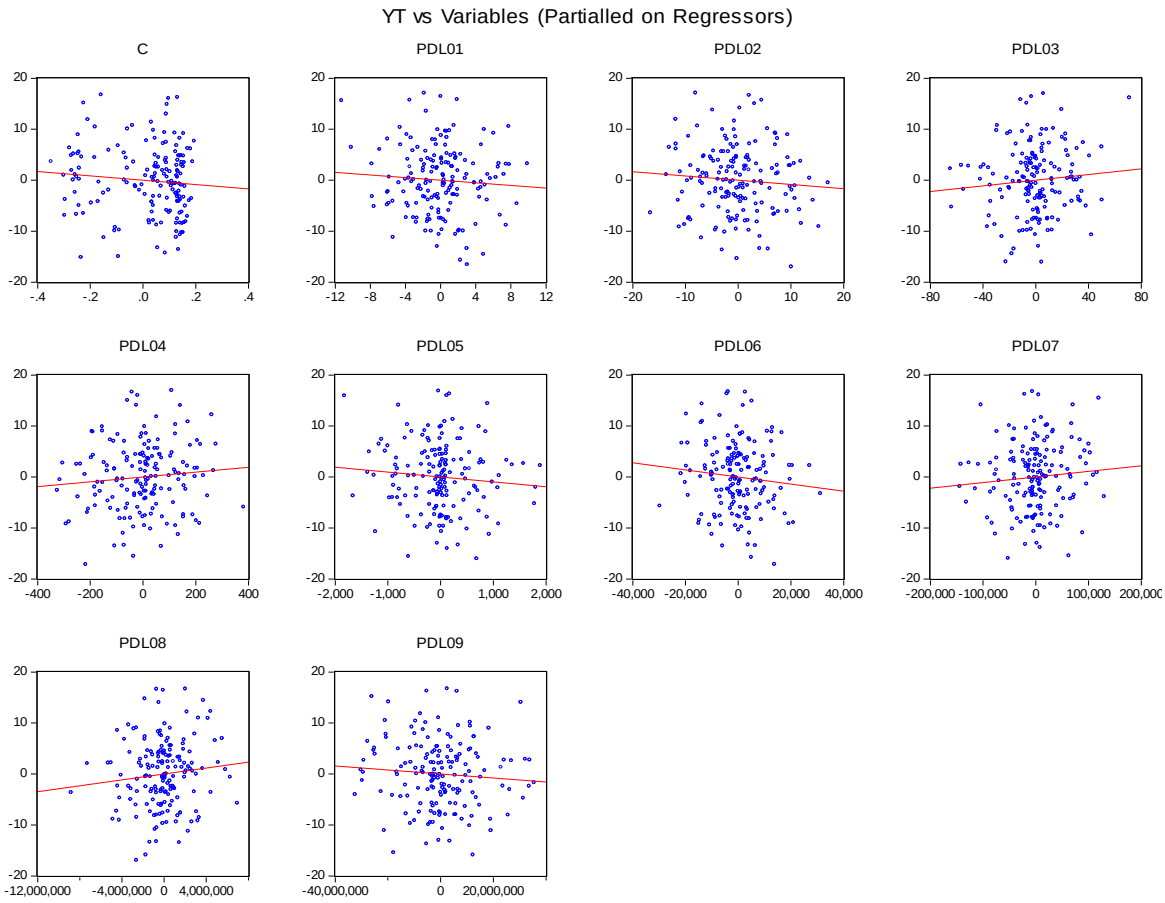


Figure 5.15: Leverage plots for the stability of diagnostics of the best PDL(26,8) model of lung cancer cases per month from 1994 to 2009.

5.14. Autoregressive Polynomial Distributed Lag (ARPDL) Models

In order to minimize the error as much as possible, we have decided to continue with the above model by using the autoregressive polynomial distributed lag (ARPDL) model, which is more flexible and parsimonious. We then continue with Model II in section 5.15. We denote this by ARPDL(p,q,k,r), where p is the length of Y_t lag, q is the degree of the polynomial of Y_t , k is the length of X_t lag, and r is the degree of the polynomial of X_t. The pdl command can be used in EViews to run the PDL and ARPDL models through the following steps:

Step 1: specify the name of the series (variables) one wants to estimate e.g. (Y_t, X_{1t}, X_{0t}).

Step 2: Specify the lag length one wishes to estimate regarding to the procedure outlined earlier.

Step 3: Specify the degree of the polynomial regarding to the procedure outlined earlier.

Step 4: Open EViews and go to Quick then choose estimate equation.

Step 5: Write the command to run e.g. (yt c pdl(x1t(-1),26,5)) for simple PDL

where Y_t is the dependent variable, c is the constant, pdl is the command, $X_{1t}(-1)$ is the independent variable one wishes to start from, 26 is the lag length, and 5 is the degree of the polynomial one has chosen in advanced. For a complex ARPDL model we choose the following command yt c pdl(x1t(-1),26,5) pdl(x0t(-1),26,5) pdl(yt(-1),12,3) where the procedure is the same except that we have added two variables we wanted to regress on which are X_{0t} and Y_t.

Step 6: Print the graphs and specify the test needed.

Note that the terms PDL1, PDL2, PDL3, PDL4 ,..., correspond to Z1, Z2, Z3, Z4,..., in Equation (5.23).

5.14.1. Choosing the Lag Length of Y_t from OLS

Using the procedure outlined in Maddala & Lahiri (2009, pp 526-533), the best lag length of Y_t is as shown (starred) in Table 5.10. We ran the regression 36 times using different lags of y_t, starting from lag 36 to lag 1. Then, we checked where the fit of the models deteriorates significantly.

Table 5.10: Choosing the best lag length of Y_t from ordinary least squares.

	Lag			
coefficient of	14	13	12*	11
y_{t-1}	0.23	0.23	0.19	0.24
y_{t-2}	0.17	0.18	0.18	0.20
y_{t-3}	0.06	0.06	0.06	0.12
y_{t-4}	0.04	0.03	0.02	0.01
y_{t-5}	-0.01	-0.00	0.00	-0.01
y_{t-6}	0.03	0.03	0.04	0.04
y_{t-7}	-0.07	-0.08	-0.08	-0.08
y_{t-8}	0.024	0.03	0.03	0.02
y_{t-9}	0.15	0.15	0.16	0.21
y_{t-10}	0.01	0.01	0.00	0.07
y_{t-11}	0.11	0.11	0.09	0.11
y_{t-12}	0.30	0.30	0.30	
y_{t-13}	-0.12	-0.09		
y_{t-14}	0.05			
Sum of coefficients	0.97	0.96	0.98	0.93
$\bar{R}^2$	0.478588	0.479481	0.478342	0.427623
F-statistic	12.60447	13.61282	14.67807	13.22528
AIC	6.723771	6.711298	6.703722	6.797215

12*= best model (lag)

From Table 5.10, the appropriate lag length of Y_t selected is 12. This is due to the steady increase in the sum of the coefficients until we use a lag of 12. We selected the lag of 12 also because the difference between the adjusted R-squared values of lag 12 and lag 13 is insignificant. Between their AICs, lag length of 12 has the lowest AIC value of about 6.70.

5.14.2. Choosing the Degree of the Polynomial Y_t

Here, we started with a ninth-degree polynomial and decreased it until we obtained a satisfactory fit.

Table 5.11: Choosing the degree of the polynomial.

	Equation								
	1			2*			3		
coefficient	6th order	t ratios	p-value	5th order	t ratios	p-value	4th order	t ratios	p-value
Z_{0t}	-0.023103	-0.57	0.57	-0.010804	-0.34	0.74	0.00461	0.16	0.88
Z_{1t}	0.012725	0.38	0.71	0.021368	0.75	0.46	-0.007137	-0.48	0.63
Z_{2t}	0.018442	0.97	0.33	0.010361	1.11	0.27	0.003821	0.51	0.61
Z_{3t}	-0.002521	-0.51	0.61	-0.004108	-1.12	0.27	5.16E-05	0.06	0.95
Z_{4t}	-0.00094	-0.58	0.56	-0.000171	-0.47	0.64	0.000122	0.46	0.65
Z_{5t}	6.35E-05	0.43	0.67	0.000117	1.16	0.25			
Z_{6t}	1.82E-05	0.49	0.62						
$\bar{R}^2$	0.483028			0.485298			0.484246		
σ^2	7591.889			7602.491			7662.062		

2*= best model (order of polynomial)

Therefore, the best order of the polynomial is 5 as shown (starred) in Table 5.11 above. Hence, the best model is ARPDL(12,5,26,8);

$$Y_t = \alpha + \sum_{i=1}^{26}(\gamma_0 i^0 + \gamma_1 i^1 + \gamma_2 i^2 + \gamma_3 i^3 + \cdots + \gamma_8 i^8)X_{t-i} + \sum_{i=1}^{12}(\gamma_0 i^0 + \gamma_1 i^1 + \gamma_2 i^2 + \gamma_3 i^3 + \cdots + \gamma_5 i^5)Y_{t-i} + \varepsilon_t \quad 5.25$$

Table 5.12: Results of the autoregressive polynomial distributed lag ARPDL(12,5,26,8) model.

Variable	Coefficient	t-Statistic	p-value
C	-6.43	-1.58	0.12
Z_{0t}	-0.10	-0.70	0.48
Z_{1t}	-0.11	-1.39	0.17
Z_{2t}	0.03	1.10	0.27
Z_{3t}	0.01	1.40	0.16
Z_{4t}	-0.00	-1.11	0.27
Z_{5t}	-0.00	-1.37	0.17
Z_{6t}	0.00	1.02	0.31
Z_{7t}	0.00	1.28	0.20
Z_{8t}	-0.00	-0.93	0.35
Z_{9t}	-0.13	-2.86	0.00
Z_{10t}	0.01	0.49	0.63
Z_{11t}	0.01	1.31	0.19
Z_{12t}	-0.00	-0.71	0.48
Z_{13t}	-0.00	-0.58	0.56
Z_{14t}	0.00	0.83	0.41

R-squared	0.594906	Mean dependent var	32.77108
Adjusted R-squared	0.554396	S.D. dependent var	9.408990
S.E. of regression	6.280836	Akaike info criterion	6.604302
Sum squared resid	5917.335	Schwarz criterion	6.904252
Log likelihood	-532.1571	Hannan-Quinn criter.	6.726054
F-statistic	14.68561	Durbin-Watson stat	1.949266
Prob(F-statistic)	0.000000		

Note that the created variables from Z_{0t} to Z_{8t} refer to the lag of X_{t-i} whereas the variables from Z_{9t} to Z_{14t} refer to the lag of Y_{t-i}.

From the three models obtained so far we prefer the ARPDL(12,5,26,8) model based on the lowest value of AIC and adjusted R-squared values. In addition, this model captures the seasonality pattern better than the OLS and the PDL models.

The fitted model is shown in Figure 5.16 together with residual diagnostic plots. This is followed by the distribution of the series in the histogram with a complement of standard descriptive statistics displayed along with the histogram (see Figure 5.17). The p-value (p=0.42) of the Jarque-Bera test is not less than 0.05 for a 5% significance level and hence we do not reject the null hypothesis that the model is normally distributed. Figure 5.18 shows leverage plots of the residuals. Here, we can see that the residuals are not collinear. Hence, we forecast with this model and present the k-step ahead forecasts.

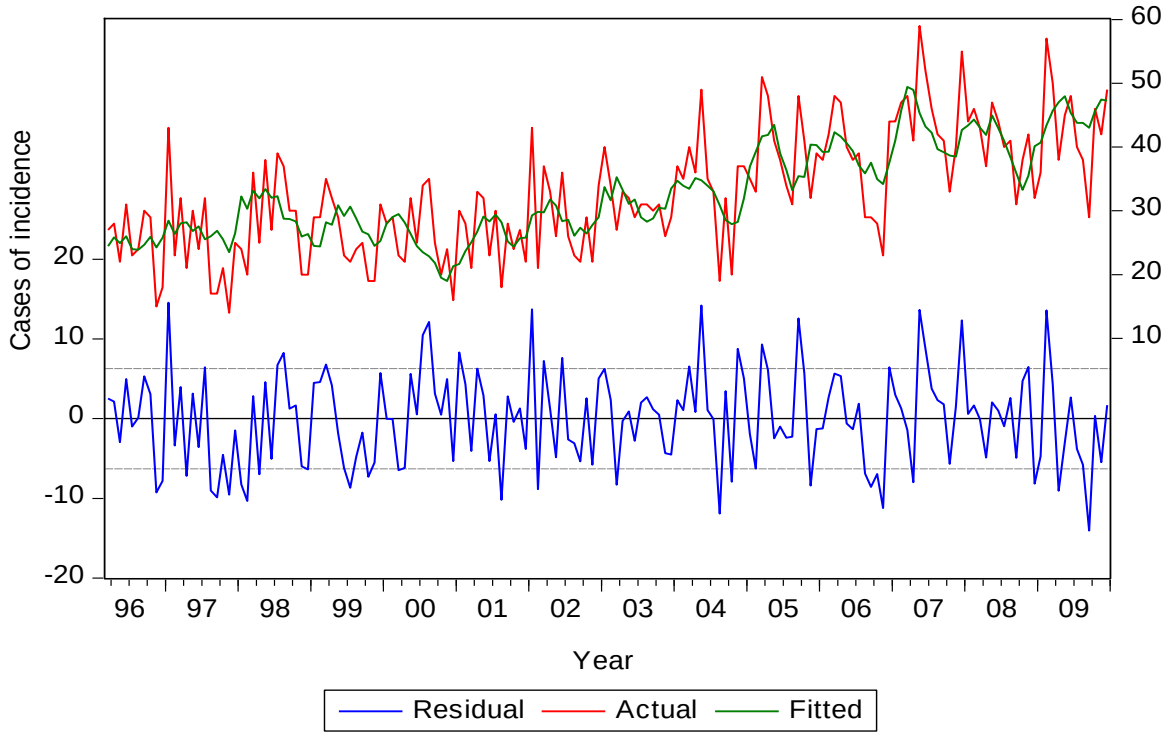


Figure 5.16: Fitted and residual plots for the best ARPDL(12,5,26,8) model of lung cancer cases per month from 1994 to 2009.

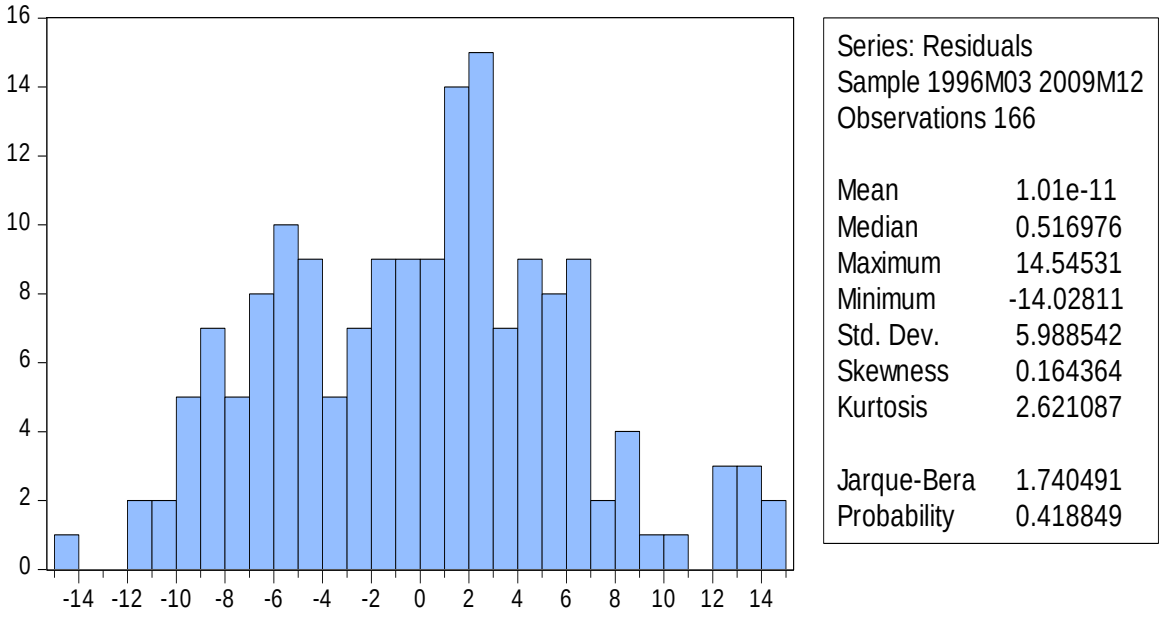


Figure 5.17: Residual diagnostic of the normality test of the best ARPDL(12,5,26,8) model of lung cancer cases per month from 1994 to 2009.

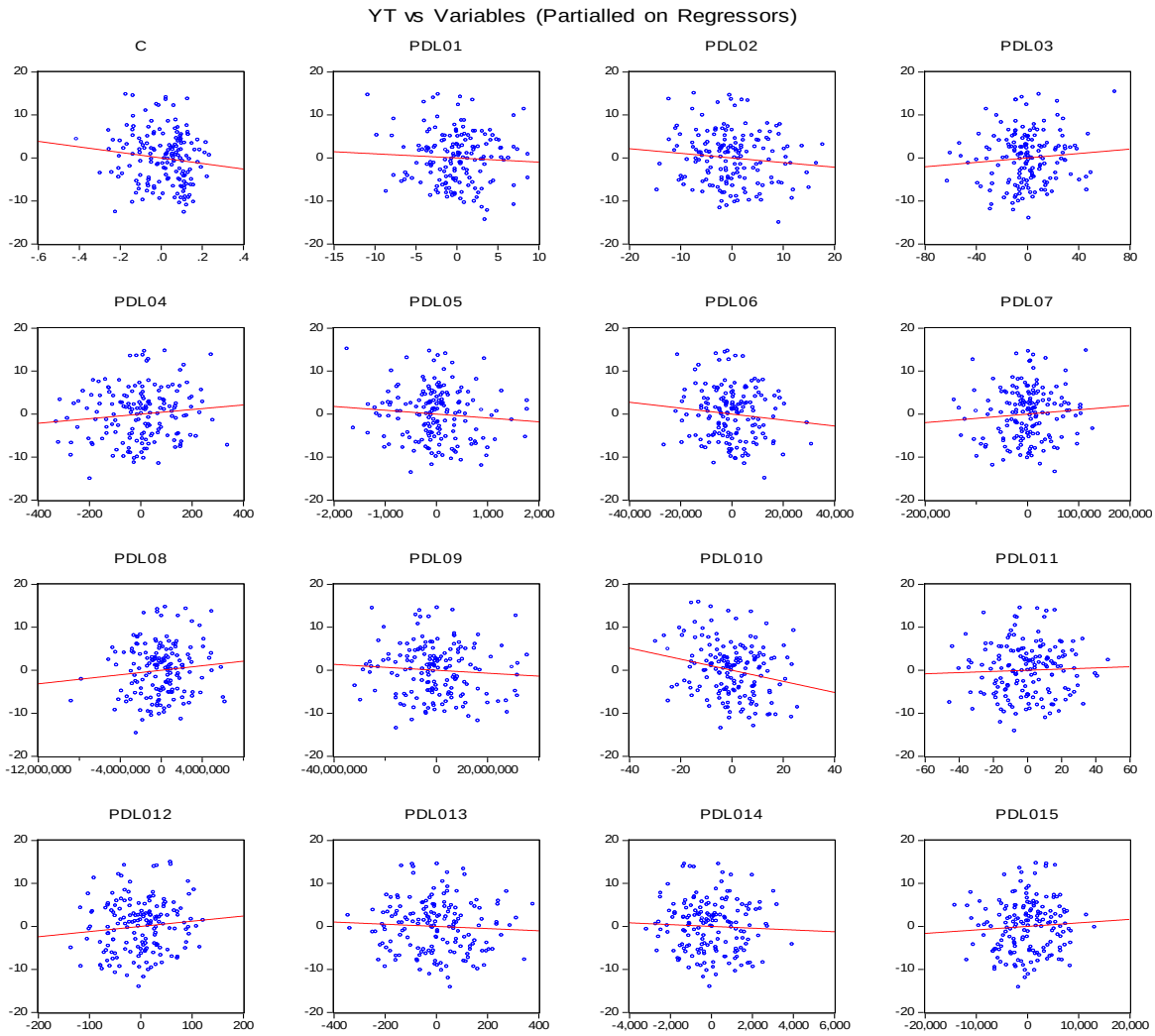


Figure 5.18: Leverage plots for the stability of diagnostics of the best ARPDL(12,5,26, 8) model of lung cancer cases per month from 1994 to 2009.

5.14.3. The Breusch-Godfrey Test for Serial Correlation

When the regression model includes lagged dependent variables as explanatory variables, the Durbin and Watson test is not valid anymore. Thus, Breusch (1978) and Godfrey (1978) developed the Lagrange Multiplier (LM) test that is applicable when a lagged dependent variable is present. Moreover, it takes into account a higher order of serial correlation. However, Durbin in 1970 developed a new test based on a *h*-statistic that can be used instead in the presence of lagged dependent variables.

Table 5.13: Results of Breusch-Godfrey LM test of ARPDL(12,5,26,8) model.

F-statistic	1.277	Prob. F(1,149)	0.26
Obs*R-squared	1.411	Prob. Chi-Square(1)	0.23

From Table 5.13, the values of both the LM-statistic and the F-statistic are low, indicating that we do not reject the null hypothesis and hence conclude that there is no significant serial correlation. Residuals generated from the model are not serially correlated because the p-values are not very small i.e. they are not less than 0.05 for a 5% significance level. For the full results see Table A13 in Appendix A.

5.14.4. Cross-validation

Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. Thus, to make sure that the results obtained so far are close to the real values, we perform cross-validation of the model by using the one step ahead out-of-sample forecasts. We fit the model from 1994 to 2007 and forecast over the period from 2008 to 2009 using the one step ahead out-of-sample forecasts and then compare with the actual observation. The forecasting performance of the fitted ARPDL(12,5,26,8) model can be seen from Figure 5.19 as the two graphs are roughly close to each other. The forecast graph appears to smooth the actual observations well. In addition, we run a new model with high number of lags using data with different period from 2000 to 2007 to assess the validity of the ARPDL model. Appendix S shows the steps of analyzing and choosing the best-fit model for this data. Therefore, the best-fit model selected is the ARPDL(11,2,23,6) model. Moreover, the one step ahead out-of-sample forecast was performed to check the performance of the model. The forecast of the ARPDL(11,2,23,6) model again seems to fit the data well and can capture the seasonal component (see Figure 5.20). However, the use of high number of lags might be a case of over-fitting. Therefore we consider a new model with fewer number of lags. This will provide us with a yardstick to compare the models with high number of lags (see Appendix G for the results). From the results obtained, the best-fit model selected is the ARPDL(3,1,6,2). Moreover, the one step ahead out-of-sample forecast was performed to check the performance of the model. Figure 5.21 illustrates the one step ahead out-of-sample forecasts from 2008 to 2009 with the actual cases. The forecast graph in this case fails to capture any seasonality in the series. The short-term forecast the reduced model eventually goes to be straight line and poor at predicting series with seasonality. Hence, we continue to present the short-term forecast for the ARPDL(12,5,26,8) model as shown in Figure 5.23.

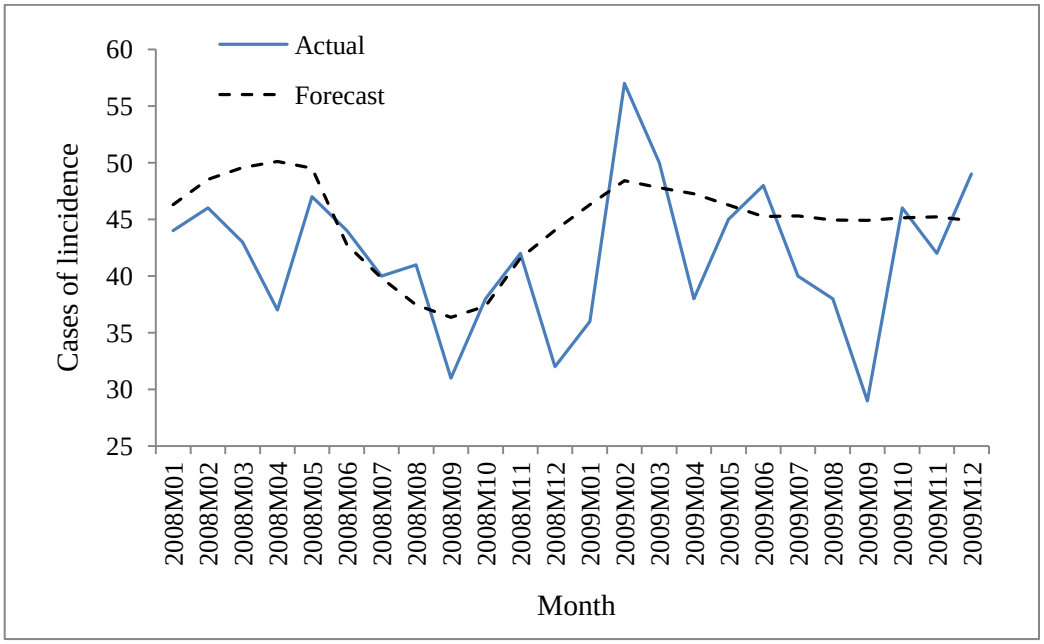


Figure 5.19: Actual and forecast ARPDL(12,5,26,8) model with 24 months ahead forecast of lung cancer cases per month from 2008 to 2009.

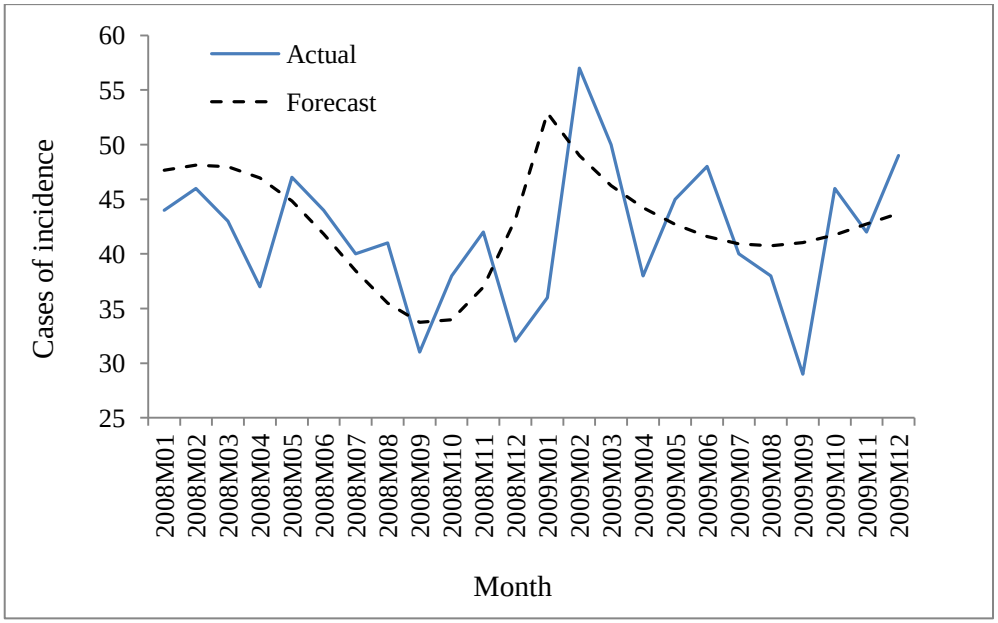


Figure 5.20: Actual and forecast ARPDL(11,2,23,6) model with 24 months ahead forecast of lung cancer cases per month from 2008 to 2009.

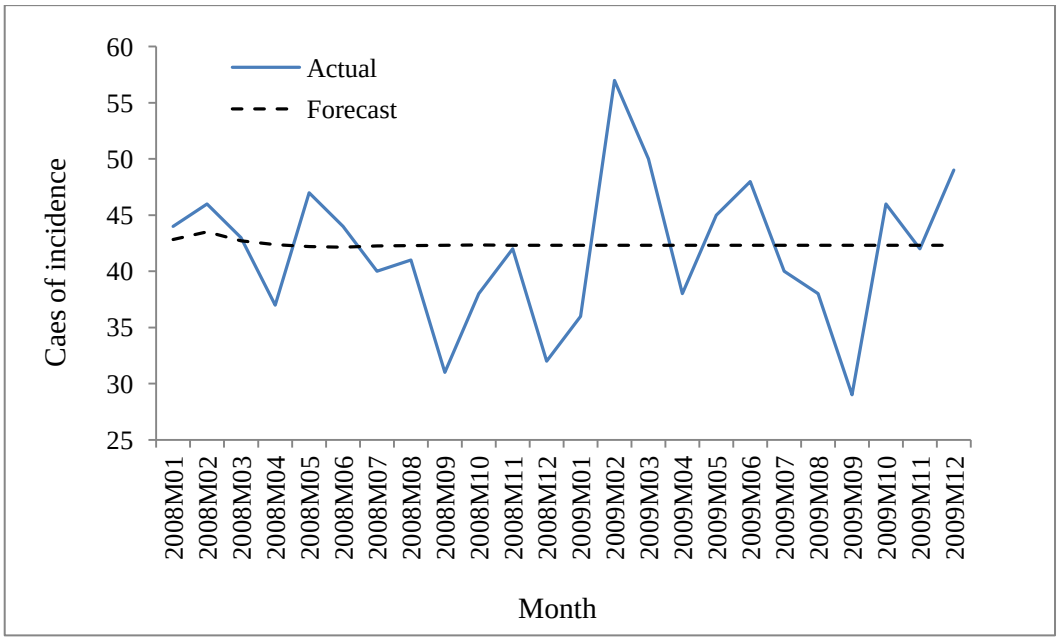


Figure 5.21: Actual and forecast ARPDL(3,1,6,2) model with 24 months ahead forecast of lung cancer cases per month from 2008 to 2009.

5.14.5. Results of the Best ARPDL(12, 5, 26,8) Model

The forecast between 2010 and 2012 of the best ARPDL(12,5,26,8) model is shown in Figure 5.22 and Figure 5.23.

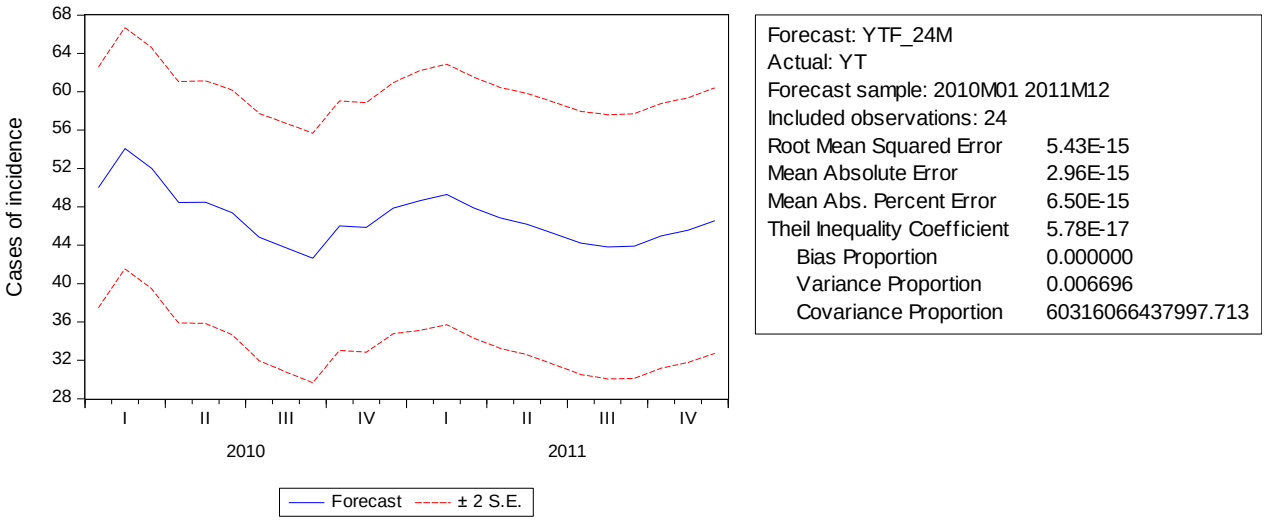


Figure 5.22: Forecast of the best ARPDL(12,5,26,8) model of lung cancer cases per month from 2010 to 2012.

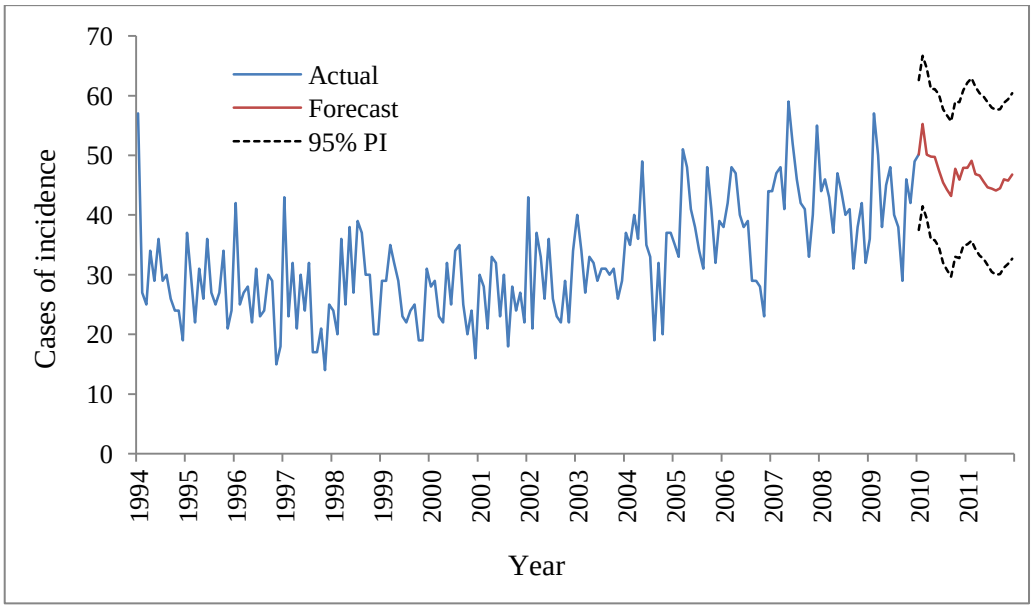


Figure 5.23: Actual and fitted ARPDL(12,5,26,8) model with 24 months ahead forecast of lung cancer cases per month from 1994 to 2012.

5.15 Model II: Dynamic Regression of Total Cases of Lung Cancer on Smoking Population Separately for Males and Females

In this section, we were seeking to find the relationship between the total cases of lung cancer from 1994 to 2009 and smoking population separately for males & females per month. Therefore, the relationship between the total cases and 36 lagged periods of male and female smoking population of the unrestricted model is

$$Y_t = \alpha + \beta_1 X_{1t-1} + \beta_2 X_{1t-2} + \cdots + \beta_{36} X_{1t-36} + \beta_1 X_{0t-1} + \beta_2 X_{0t-2} + \cdots + \beta_{36} X_{0t-36} + \varepsilon_t \quad 5.26$$

where

Y_t = the incidence at time t (number of cases in month t).

Using EViews8 software package, the results are presented as in Davidson & MacKinnon (1993) and Maddala & Lahiri (2009, pp 526-533) are shown in Table 5.14. We ran the regression 36 times using different lags, starting from lag 36 to lag 1. Then, we checked where the fit of the models deteriorates significantly.

Table 5.14: Choosing the lag length from OLS.

Coefficient	lag		
	27	26*	25
x_{1t-1}	-4.07	-2.70	-2.44
x_{1t-2}	7.35	4.66	4.44
x_{1t-3}	-6.56	-5.18	-5.18
x_{1t-4}	4.87	4.87	4.87
x_{1t-5}	-2.00	-2.00	-2.00
x_{1t-6}	-1.73	-1.73	-1.74
x_{1t-7}	2.41	2.41	2.41
x_{1t-8}	1.91	1.91	1.92
x_{1t-9}	-2.93	-2.92	-2.93
x_{1t-10}	0.06	0.06	0.06
x_{1t-11}	6.16	6.16	6.22
x_{1t-12}	-11.35	-11.36	-8.51
x_{1t-13}	4.83	6.70	0.80
x_{1t-14}	3.52	-0.34	2.70
x_{1t-15}	-3.90	-1.82	-1.82
x_{1t-16}	3.03	3.03	3.03
x_{1t-17}	-2.78	-2.78	-2.78
x_{1t-18}	-1.31	-1.31	-1.31
x_{1t-19}	2.41	2.41	2.41
x_{1t-20}	2.21	2.21	2.21
x_{1t-21}	-3.15	-3.15	-3.15
x_{1t-22}	1.28	1.27	1.27
x_{1t-23}	5.47	5.48	5.58
x_{1t-24}	-15.48	-15.50	-11.24
x_{1t-25}	13.46	14.35	5.17
x_{1t-26}	-2.78	-4.84	-
x_{1t-27}	-1.20	-	-
x_{0t-1}	2.41	-0.47	-1.62
x_{0t-2}	-7.74	-2.08	-1.03
x_{0t-3}	23.14	20.21	20.21
x_{0t-4}	-25.72	-25.73	-25.73
x_{0t-5}	12.71	12.71	12.70
x_{0t-6}	-3.21	-3.20	-3.20
x_{0t-7}	-6.44	-6.44	-6.44
x_{0t-8}	0.51	0.51	0.50
x_{0t-9}	15.63	15.64	15.64
x_{0t-10}	-24.80	-24.81	-24.81
x_{0t-11}	12.53	12.52	12.43
x_{0t-12}	12.12	12.15	12.33
x_{0t-13}	-8.09	-15.37	-16.52
x_{0t-14}	-8.85	6.28	7.18
x_{0t-15}	19.87	11.75	11.75
x_{0t-16}	-23.42	-23.42	-23.42
x_{0t-17}	16.49	16.49	16.48
x_{0t-18}	-6.66	-6.66	-6.66
x_{0t-19}	0.40	0.40	0.40
x_{0t-20}	-3.05	-3.05	-3.05
x_{0t-21}	12.47	12.48	12.48
x_{0t-22}	-24.35	-24.35	-24.35
x_{0t-23}	15.58	15.58	15.56
x_{0t-24}	5.55	5.56	7.46
x_{0t-25}	3.14	-3.58	-7.67
x_{0t-26}	-16.38	-2.09	-
x_{0t-27}	7.78	-	-
Sum of coefficients	1.376276	0.913401	0.624424
$\bar{R}^2$	0.494108	0.497934	0.484655

The adjusted R-squared increased gradually until 26 lags. As a result, it appears that a lag distribution using 26 lags is appropriate and this agrees with our previous analysis when the data is total smoking population (Model I). Table 5.15 below shows the DW test for different lengths of the lag distribution:

Table 5.15: The Durbin-Watson statistic.

Length of lag	DW
28	1.49
27	1.50
26	1.50
25	1.52

From Table 5.15, this suggests a typical symptom of collinearity and we should be estimating some more general dynamic models, allowing for autocorrelated errors.

Table 5.16: The best-unrestricted least squares (OLS) model with 26 lags.

Variable	Coefficient	t-Statistic	p-value
C	12.56	0.61	0.54
x_{1t-1}	-2.70	-1.45	0.15
x_{1t-2}	4.66	1.23	0.22
x_{1t-3}	-5.18	-1.27	0.20
x_{1t-4}	4.86	1.20	0.23
x_{1t-5}	-2.00	-0.49	0.62
x_{1t-6}	-1.73	-0.42	0.67
x_{1t-7}	2.41	0.59	0.55
x_{1t-8}	1.91	0.47	0.64
x_{1t-9}	-2.92	-0.72	0.47
x_{1t-10}	0.06	0.01	0.99
x_{1t-11}	6.16	1.50	0.14
x_{1t-12}	-11.35	-2.63	0.01
x_{1t-13}	6.69	1.51	0.13
x_{1t-14}	-0.34	-0.08	0.94
x_{1t-15}	-1.82	-0.43	0.66
x_{1t-16}	3.02	0.72	0.47
x_{1t-17}	-2.78	-0.66	0.50
x_{1t-18}	-1.31	-0.31	0.75
x_{1t-19}	2.41	0.57	0.56
x_{1t-20}	2.21	0.53	0.60
x_{1t-21}	-3.15	-0.75	0.45
x_{1t-22}	1.27	0.30	0.76
x_{1t-23}	5.48	1.25	0.21
x_{1t-24}	-15.50	-3.08	0.00
x_{1t-25}	14.35	3.01	0.00
x_{1t-26}	-4.84	-2.15	0.03
x_{0t-1}	-0.47	-0.09	0.93
x_{0t-2}	-2.08	-0.19	0.84
x_{0t-3}	20.21	1.79	0.08
x_{0t-4}	-25.73	-2.27	0.02
x_{0t-5}	12.70	1.12	0.26
x_{0t-6}	-3.20	-0.28	0.78

Table 5.16 Continued.

x_{0t-7}	-6.44	-0.57	0.57
x_{0t-8}	0.50	0.04	0.96
x_{0t-9}	15.63	1.38	0.17
x_{0t-10}	-24.80	-2.19	0.03
x_{0t-11}	12.52	1.10	0.27
x_{0t-12}	12.14	1.04	0.29
x_{0t-13}	-15.36	-1.28	0.20
x_{0t-14}	6.28	0.54	0.59
x_{0t-15}	11.74	1.05	0.29
x_{0t-16}	-23.41	-2.10	0.04
x_{0t-17}	16.48	1.48	0.14
x_{0t-18}	-6.65	-0.59	0.55
x_{0t-19}	0.39	0.03	0.97
x_{0t-20}	-3.05	-0.27	0.78
x_{0t-21}	12.47	1.12	0.26
x_{0t-22}	-24.34	-2.19	0.03
x_{0t-23}	15.58	1.40	0.16
x_{0t-24}	5.56	0.49	0.61
x_{0t-25}	-3.57	-0.34	0.73
x_{0t-26}	-2.08	-0.41	0.68

R-squared	0.656161	Mean dependent var	32.77108
Adjusted R-squared	0.497934	S.D. dependent var	9.408990
S.E. of regression	6.666895	Akaike info criterion	6.886140
Sum squared resid	5022.567	Schwarz criterion	7.879726
Log likelihood	-518.5496	Hannan-Quinn criter.	7.289443
F-statistic	4.146956	Durbin-Watson stat	1.503597
Prob(F-statistic)	0.000000		

The fitted model is shown in Figure 5.24 together with residual diagnostic plots. This is followed by the distribution of the series in the histogram with a complement of standard descriptive statistics displayed along with the histogram (see Figure 5.25). The p-value (p=0.33) of the Jarque-Bera test is not less than 0.05 for a 5% significance level and hence we do not reject the null hypothesis that the model is normally distributed. Figure L1 in Appendix L shows leverage plots of the residuals. As we can see from Figure L1, the points are compressed towards the vertical line indicating collinearity between the terms. Therefore we look for an adequate model that is more flexible and parsimonious.

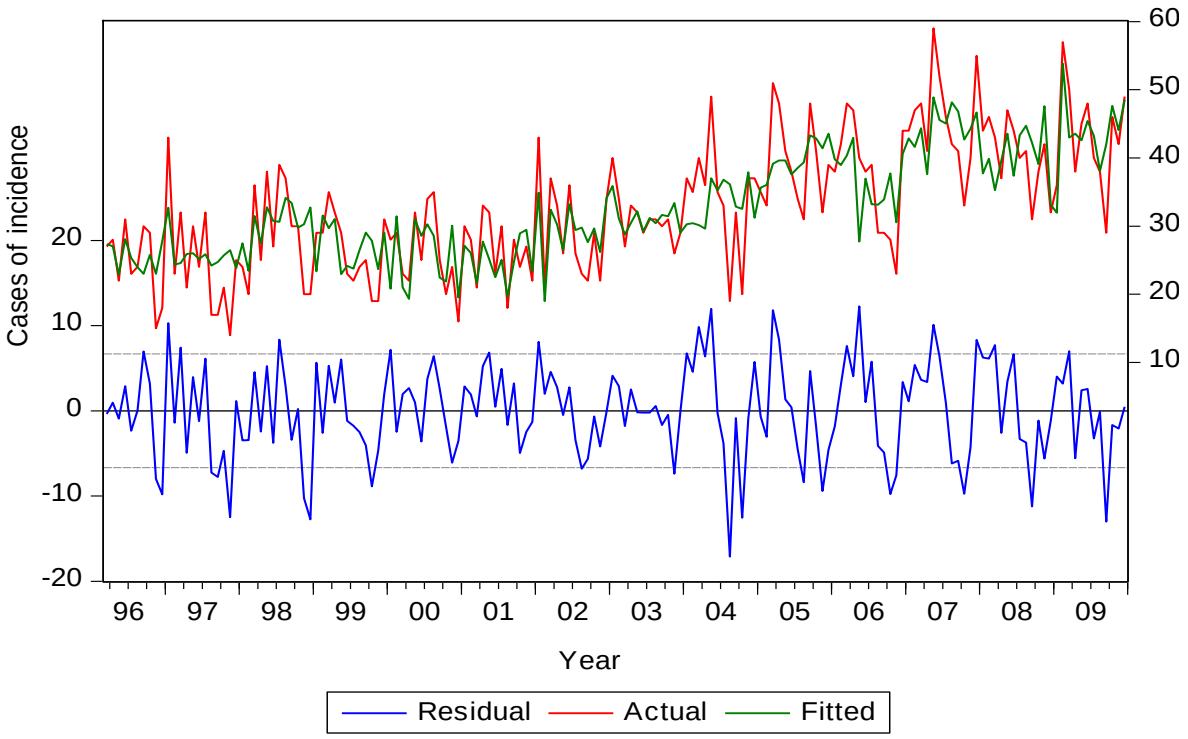


Figure 5.24: Fitted and residual plots for the best OLS model of lung cancer cases per month from 1994 to 2009.

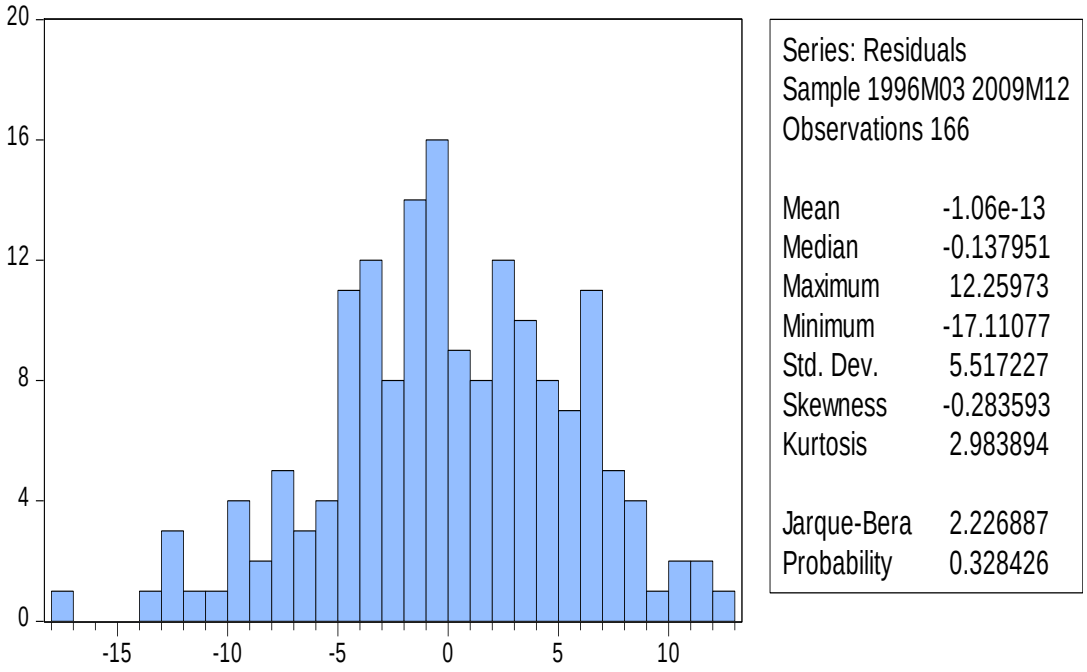


Figure 5.25: Normality plot of the best OLS model of lung cancer cases per month from 1994 to 2009.

5.15.1. Choosing the Degree of the Polynomial

Starting with a high-degree of polynomial, we obtain the following table.

Table 5.17: Choosing the degree of the polynomial.

Coefficient	Equation								
	9th order	t ratios	p-value	8th order*	t ratios	p-value	7th order	t ratios	p-value
Z_{0t}	-0.320899	-1.65	0.10	-0.348903	-1.90	0.06	-0.370668	-2.09	0.04
Z_{1t}	-0.082038	-0.52	0.60	-0.020377	-0.20	0.84	-0.037517	-0.37	0.71
Z_{2t}	0.033154	0.91	0.36	0.04173	1.35	0.18	0.042385	1.82	0.07
Z_{3t}	0.008062	0.58	0.57	0.001177	0.23	0.82	0.001986	0.42	0.68
Z_{4t}	-0.000318	-0.23	0.82	-0.000713	-0.67	0.51	-0.000663	-1.51	0.13
Z_{5t}	-0.000214	-0.57	0.57	-1.86E-05	-0.27	0.78	-2.77E-05	-0.47	0.64
Z_{6t}	-1.97E-06	-0.11	0.91	3.54E-06	0.27	0.79	2.53E-06	1.24	0.22
Z_{7t}	2.04E-06	0.55	0.58	8.05E-08	0.29	0.78	1.08E-07	0.50	0.62
Z_{8t}	1.86E-08	0.26	0.79	-4.34E-09	-0.09	0.93	1.038911	2.08	0.04
Z_{9t}	-6.31E-09	-0.53	0.60	0.684081	1.28	0.20	-0.251474	-0.87	0.38
Z_{10t}	0.699939	1.23	0.22	-0.385987	-1.31	0.19	-0.118809	-1.80	0.07
Z_{11t}	-0.449985	-0.96	0.34	0.003509	0.04	0.97	0.01107	0.80	0.42
Z_{12t}	-0.004602	-0.04	0.97	0.022261	1.47	0.14	0.001965	1.56	0.12
Z_{13t}	0.029514	0.70	0.49	-0.003479	-1.09	0.28	-0.000119	-0.68	0.50
Z_{14t}	-0.003042	-0.70	0.48	-0.000326	-1.55	0.12	-7.99E-06	-1.35	0.18
Z_{15t}	-0.000534	-0.47	0.64	6.15E-05	1.62	0.11	3.70E-07	0.56	0.58
Z_{16t}	5.50E-05	1.00	0.32	1.38E-06	1.58	0.12			
Z_{17t}	3.48E-06	0.31	0.76	-2.61E-07	-1.84	0.06			
Z_{18t}	-2.33E-07	-1.08	0.28						
Z_{19t}	-6.79E-09	-0.19	0.85						
$\bar{R}^2$	0.498237			0.503602			0.496625		
σ^2	6440.995			6460.014			6639.943		
DW	1.829387			1.831135			1.812622		

First, we test the coefficients of Z_{9t} and Z_{19t} for males and females respectively at the 5% level and we do not reject the hypotheses that they are zero (p-value 0.60 for males and 0.85 for females). Next, we test the coefficients of Z_{8t} and Z_{17t} for males and females respectively; also, we do not reject the hypotheses that their coefficients are zero (p-value 0.93 for males and 0.06 for females). We then compare the adjusted R-squared values for the three models and their corresponding DW statistics to select the best order for the polynomial. From Table 5.17 above, the eighth-order polynomial was chosen because of its highest adjusted R-squared and DW statistic. Hence, the model as illustrated in Table 5.18.

Table 5.18: Results of restricted least squared PDL(26,8) model.

Variable	Coefficient	t-Statistic	p-value
C	3.94	0.19	0.84
Z_{0t}	-0.34	-1.89	0.06
Z_{1t}	-0.02	-0.19	0.84
Z_{2t}	0.04	1.35	0.18
Z_{3t}	0.00	0.23	0.82
Z_{4t}	-0.00	-0.66	0.51
Z_{5t}	-0.00	-0.27	0.78
Z_{6t}	0.00	0.26	0.79
Z_{7t}	0.00	0.28	0.78
Z_{8t}	-0.00	-0.08	0.93
Z_{9t}	0.68	1.28	0.20
Z_{10t}	-0.38	-1.30	0.19
Z_{11t}	0.00	0.03	0.97
Z_{12t}	0.02	1.47	0.14
Z_{13t}	-0.00	-1.08	0.28
Z_{14t}	-0.00	-1.55	0.12
Z_{15t}	0.00	1.61	0.11
Z_{16t}	0.00	1.57	0.12
Z_{17t}	-0.00	-1.83	0.06

R-squared	0.557754	Mean dependent var	32.77108
Adjusted R-squared	0.503602	S.D. dependent var	9.408990
S.E. of regression	6.629154	Akaike info criterion	6.728192
Sum squared resid	6460.014	Schwarz criterion	7.084383
Log likelihood	-539.4399	Hannan-Quinn criter.	6.872772
F-statistic	10.29969	Durbin-Watson stat	1.831135
Prob(F-statistic)	0.000000		

Note that the created variables from Z_{0t} to Z_{8t} refer to the lag of X_{1t-i} whereas the variables from Z_{9t} to Z_{17t} refer to the lag of X_{0t-i}. What the polynomial approximation has done is to reduce the number of parameters that have to be estimated from 52 to just 18 in the restricted equation. Therefore, the procedure reduced any multicollinearity problems that might arise in Equation (5.26).

The fitted model is shown in Figure 5.26 together with residual diagnostic plots. This is followed by the distribution of the series in the histogram with a complement of standard descriptive statistics displayed along with the histogram (see Figure 5.27). The p-value (p=0.97) of the Jarque-Bera test is not less than 0.05 for a 5% significance level and hence we do not reject the null hypothesis that the model is normally distributed. Figure L2 in Appendix L shows leverage plots of the residuals. We can see that the residuals are not collinear but the fitted model does not clearly reflect the seasonal nature of the data. Therefore, we fit a new ARPDL model in the next section.

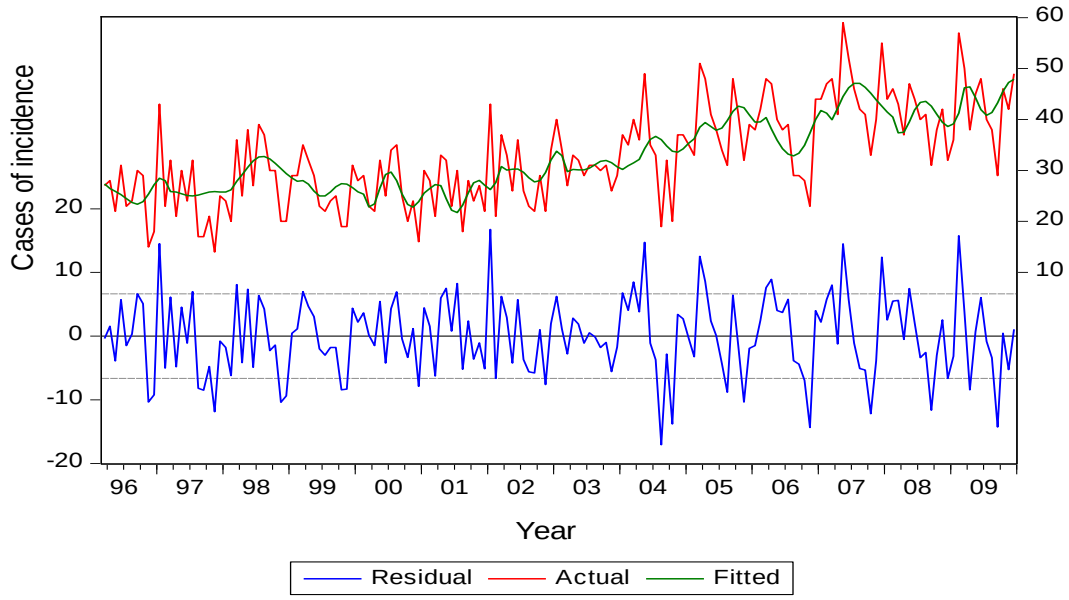


Figure 5.26: Fitted and residual plots for the best PDL(26,8) model of lung cancer cases per month from 1994 to 2009.

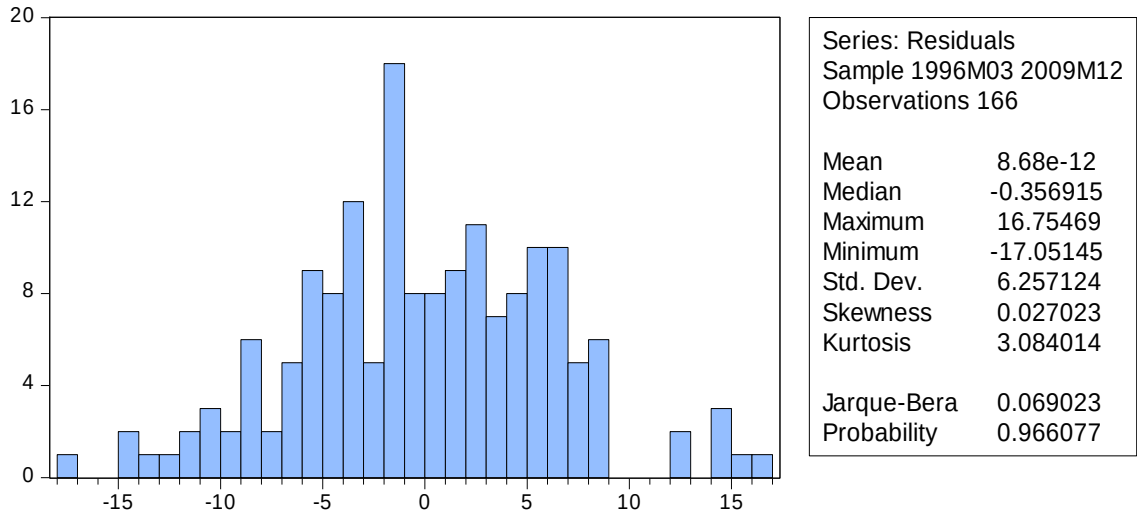


Figure 5.27: Normality plot of the best PDL(26,8) model of lung cancer cases per month from 1994 to 2009.

5.16. Autoregressive Polynomial Distributed Lagged (ARPDL) Variables

Using the procedure outlined in Maddala & Lahiri (2009, pp 526-533), the best lag length of Y_t was 12 lags.

5.16.1. Choosing the Degree of the Polynomial of Y_t

Here, we started with a ninth-degree polynomial and decreased it until we obtained a satisfactory fit.

Table 5.19: Choosing the degree of the polynomial.

Coefficient	Equation								
	5th order	t ratios	p-value	4th order	t ratios	p-value	3th order*	t ratios	p-value
Z_{0t}	-0.385635	-2.33	0.02	-0.388004	-2.35	0.02	-0.383201	-2.32	0.02
Z_{1t}	-0.155767	-1.59	0.11	-0.15897	-1.63	0.11	-0.163728	-1.68	0.09
Z_{2t}	0.046728	1.72	0.09	0.047123	1.74	0.08	0.047303	1.75	0.08
Z_{3t}	0.005648	1.21	0.23	0.005756	1.24	0.22	0.006014	1.30	0.20
Z_{4t}	-0.000713	-0.76	0.45	-0.000723	-0.77	0.44	-0.000763	-0.82	0.41
Z_{5t}	-5.09E-05	-0.83	0.41	-5.17E-05	-0.84	0.40	-5.54E-05	-0.91	0.36
Z_{6t}	2.13E-06	0.18	0.85	2.21E-06	0.19	0.85	2.91E-06	0.25	0.80
Z_{7t}	1.25E-07	0.50	0.62	1.26E-07	0.51	0.61	1.42E-07	0.57	0.57
Z_{8t}	4.02E-09	0.09	0.93	3.84E-09	0.09	0.93	8.36E-10	0.02	0.98
Z_{9t}	0.938096	1.96	0.05	0.942487	1.97	0.05	0.96844	2.03	0.04
Z_{10t}	-0.192688	-0.70	0.49	-0.192801	-0.70	0.49	-0.173853	-0.64	0.53
Z_{11t}	-0.002234	-0.03	0.98	-0.002571	-0.03	0.98	-0.012249	-0.15	0.88
Z_{12t}	0.017986	1.28	0.20	0.018267	1.30	0.19	0.017162	1.24	0.22
Z_{13t}	-0.004206	-1.43	0.15	-0.004206	-1.43	0.15	-0.00377	-1.32	0.19
Z_{14t}	-0.000329	-1.68	0.09	-0.000336	-1.72	0.09	-0.000319	-1.65	0.10
Z_{15t}	7.57E-05	2.14	0.03	7.59E-05	2.14	0.03	7.03E-05	2.04	0.04
Z_{16t}	1.57E-06	1.92	0.06	1.60E-06	1.97	0.05	1.52E-06	1.90	0.06
Z_{17t}	-3.23E-07	-2.44	0.02	-3.24E-07	-2.45	0.02	-3.03E-07	-2.35	0.02
Z_{18t}	-0.182605	-3.55	0.00	-0.174873	-3.48	0.00	-0.161959	-3.48	0.00
Z_{19t}	0.001594	0.06	0.95	-0.014458	-1.00	0.32	-0.009276	-0.75	0.45
Z_{20t}	0.015134	1.76	0.08	0.011488	1.65	0.10	0.006789	3.94	0.00
Z_{21t}	-0.000735	-0.22	0.83	0.001615	1.91	0.06	0.001246	1.90	0.06
Z_{22t}	-0.000339	-1.00	0.32	-0.000175	-0.70	0.49			
Z_{23t}	6.62E-05	0.73	0.47						
$\bar{R}^2$	0.6235			0.6247			0.6261		
σ^2	4700.01			4717.6210			4733.7240		

From Table 5.19 above, the best order of the polynomial is 3 with the highest adjusted R-squared. Hence, the best model of the ARPDL(12,3,26,8) is as in the following equation:

$$Y_t = \alpha + \sum_{i=1}^{26}(\gamma_0 i^0 + \gamma_1 i^1 + \gamma_2 i^2 + \gamma_3 i^3 + \cdots + \gamma_8 i^8)X_{1t-i} + \sum_{i=1}^{26}(\gamma_0 i^0 + \gamma_1 i^1 + \gamma_2 i^2 + \gamma_3 i^3 + \cdots + \gamma_8 i^8)X_{0t-i} + \sum_{i=1}^{12}(\gamma_0 i^0 + \gamma_1 i^1 + \gamma_2 i^2 + \gamma_3 i^3)Y_{t-i} + \varepsilon_t \quad 5.27$$

Therefore, the results of ARPDL(12,3,26,8) are shown in Table 5.20 below.

Table 5.20: Results of the autoregressive polynomial distributed lag ARPDL(12,3,26,8) model.

Variable	Coefficient	Std. Error	t-Statistic	p-value
C	20.85	18.16	1.14	0.25
Z_{0t}	-0.38	0.16	-2.32	0.02
Z_{1t}	-0.16	0.09	-1.68	0.09
Z_{2t}	0.05	0.02	1.75	0.08
Z_{3t}	0.01	0.00	1.30	0.20
Z_{4t}	-0.00	0.00	-0.82	0.41
Z_{5t}	-0.00	0.00	-0.91	0.36
Z_{6t}	0.00	0.00	0.25	0.80
Z_{7t}	0.00	0.00	0.57	0.57
Z_{8t}	0.00	0.00	0.02	0.98
Z_{9t}	0.97	0.47	2.03	0.04
Z_{10t}	-0.17	0.27	-0.64	0.53
Z_{11t}	-0.01	0.08	-0.15	0.88
Z_{12t}	0.02	0.01	1.23	0.22
Z_{13t}	-0.00	0.00	-1.32	0.19
Z_{14t}	-0.00	0.00	-1.65	0.10
Z_{15t}	0.00	0.00	2.04	0.04
Z_{16t}	0.00	0.00	1.90	0.06
Z_{17t}	-0.00	0.00	-2.35	0.02
Z_{18t}	-0.16	0.04	-3.48	0.00
Z_{19t}	-0.01	0.01	-0.75	0.45
Z_{20t}	0.01	0.00	3.94	0.00
Z_{21t}	0.00	0.00	1.90	0.06

R-squared	0.675934	Mean dependent var	32.77108
Adjusted R-squared	0.626078	S.D. dependent var	9.408990
S.E. of regression	5.753518	Akaike info criterion	6.465465
Sum squared resid	4733.724	Schwarz criterion	6.896644
Log likelihood	-513.6336	Hannan-Quinn criter.	6.640484
F-statistic	13.55767	Durbin-Watson stat	2.028012
Prob(F-statistic)	0.000000		

Lag Distribution of x_{1t-i}	i	Coefficient	Std. Error	t-Statistic
*. |	1	-0.07	0.68	-0.11
* . |	2	-0.54	0.44	-1.22
* . |	3	-0.54	0.43	-1.25
* . |	4	-0.29	0.25	-1.14
.* |	5	0.04	0.24	0.18
. * |	6	0.34	0.25	1.35
. * |	7	0.54	0.21	2.56
. * |	8	0.59	0.18	3.31
. * |	9	0.51	0.19	2.69
. * |	10	0.32	0.20	1.61
.* |	11	0.07	0.19	0.39
* . |	12	-0.17	0.17	-1.02
* . |	13	-0.38	0.16	-2.32
* . |	14	-0.49	0.16	-3.00
* . |	15	-0.48	0.17	-2.80
* . |	16	-0.35	0.19	-1.88
* . |	17	-0.13	0.19	-0.67
. * |	18	0.13	0.18	0.74
. * |	19	0.39	0.18	2.14

Table 5.20 Continued.

. * |	20	0.55	0.22	2.44
. * |	21	0.54	0.26	2.02
. * |	22	0.34	0.25	1.36
* |	23	-0.03	0.27	-0.12
* . |	24	-0.44	0.48	-0.93
* . |	25	-0.61	0.48	-1.26
* |	26	-0.02	0.74	-0.03
Sum of Lags		-0.19833	0.27173	-0.72987

Lag Distribution of x_{0t-i}	i	Coefficient	Std. Error	t-Statistic
* . |	1	-2.14	1.94	-1.10
. * |	2	4.62	1.35	3.41
. * |	3	3.27	1.27	2.57
.* |	4	0.15	0.72	0.21
* . |	5	-2.05	0.73	-2.77
* . |	6	-2.72	0.76	-3.55
* . |	7	-2.19	0.62	-3.53
* . |	8	-1.11	0.52	-2.11
* |	9	-0.02	0.55	-0.04
.* |	10	0.73	0.56	1.29
.* |	11	1.08	0.52	2.08
.* |	12	1.10	0.47	2.32
.* |	13	0.96	0.47	2.03
.* |	14	0.79	0.48	1.63
.* |	15	0.64	0.49	1.30
.* |	16	0.46	0.51	0.91
.* |	17	0.17	0.53	0.32
* |	18	-0.31	0.52	-0.60
*. |	19	-0.97	0.49	-1.96
* . |	20	-1.58	0.58	-2.72
* . |	21	-1.75	0.71	-2.46
* . |	22	-1.01	0.68	-1.46
.* |	23	0.85	0.69	1.23
. * |	24	3.20	1.21	2.63
. * |	25	3.55	1.28	2.76
* . |	26	-3.86	1.93	-1.99
Sum of Lags		1.89335	1.18804	1.59367

Lag Distribution of y_{t-i}	i	Coefficient	Std. Error	t-Statistic
* . |	1	-0.10	0.06	-1.48
* . |	2	-0.09	0.04	-2.10
* . |	3	-0.10	0.04	-2.40
* . |	4	-0.12	0.04	-2.68
* . |	5	-0.14	0.04	-3.10
* . |	6	-0.16	0.04	-3.47
* . |	7	-0.16	0.04	-3.47
* . |	8	-0.14	0.04	-2.95
* . |	9	-0.09	0.04	-1.95
*. |	10	-0.01	0.04	-0.23
. * |	11	0.11	0.04	2.51
. * |	12	0.29	0.06	4.25
Sum of Lags		-0.73891	0.47365	-1.56005

Note that the created variables from Z_{0t} to Z_{8t} refer to the lag of X_{1t-i} and the variables from Z_{9t} to Z_{17t} refer to the lag of X_{0t-i} whereas the variables from Z_{18t} to Z_{21t} refer to the lag of Y_{t-i}.

The fitted model is shown in Figure 5.28 together with residual diagnostic plots. This is followed by the distribution of the series in the histogram with a complement of standard descriptive statistics displayed along with the histogram (see Figure 5.29). The p-value (p=0.86) of the Jarque-Bera test is not less than 0.05 for a 5% significance level and hence we do not reject the null hypothesis that the model is normally distributed. Figure L3 in Appendix L shows leverage plots of the residuals. Here, we can see that the residuals are not collinear. Hence, we forecast with this model and present the k-step ahead forecasts.

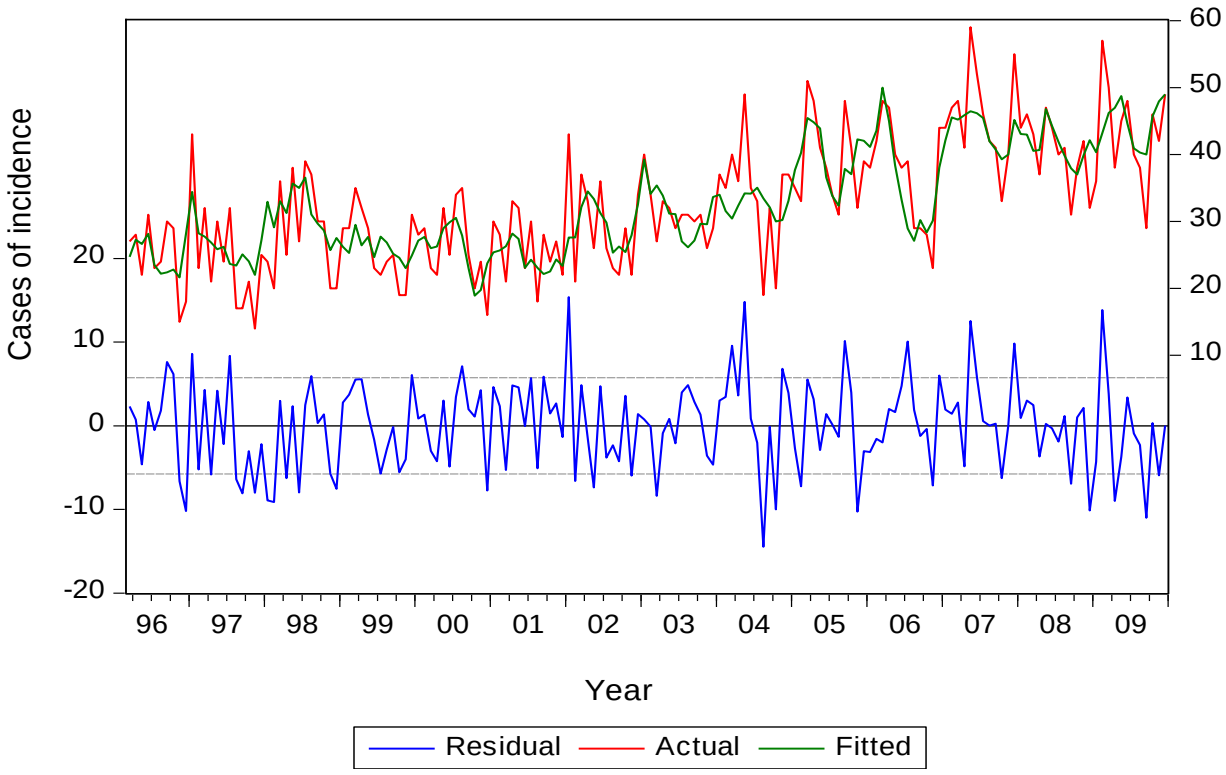


Figure 5.28: Fitted and residual plots for the best ARPDL(12,3,26,8) model of lung cancer cases per month from 1994 to 2009.

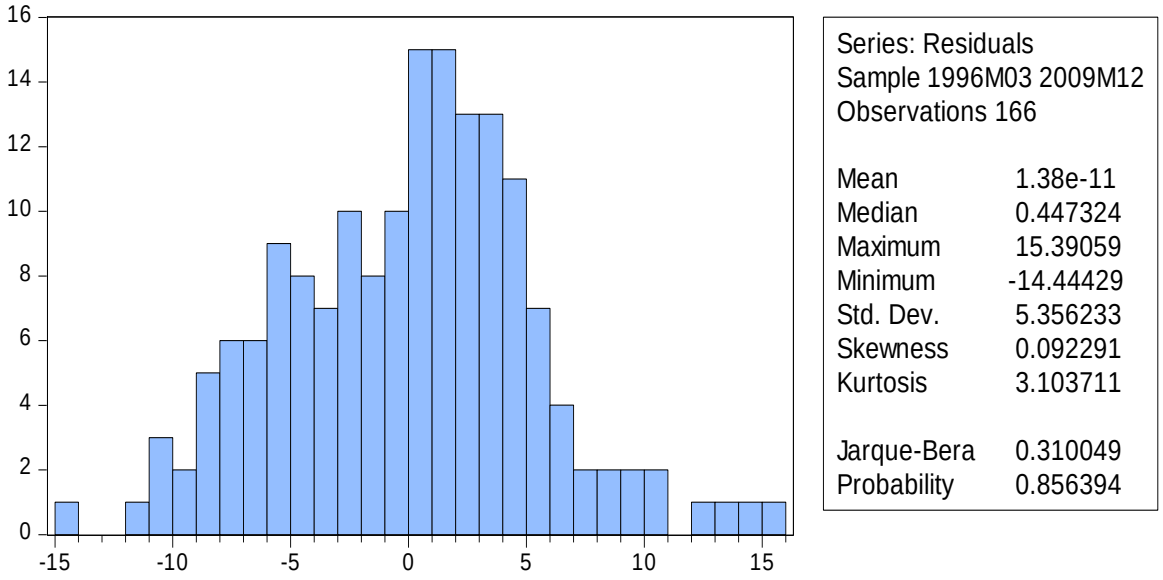


Figure 5.29: Residual diagnostic of the normality test of the best ARPDL(12,3,26,8) model of lung cancer cases per month from 1994 to 2009.

5.16.2. The Breusch-Godfrey LM Test

From Table 5.21, the values of both the LM-statistic and the F-statistic are very small, indicating that we do not reject the null hypothesis and hence conclude there is no significant serial correlation. Residuals generated from the model are not serially correlated because the p-values are not very small i.e. they are not less than 0.05 for a 5% significance level. For the full results see Table A14 in Appendix A.

Table 5.21: Results of Breusch-Godfrey LM test of ARPDL(12,3,26,8) model.

F-statistic	0.090604	Prob. F(1,142)	0.7639
Obs*R-squared	0.105850	Prob. Chi-Square(1)	0.7449

Hence, we forecast this model and present the k-step ahead forecasts as shown in Figure 5.31.

5.16.3. Results of the Best ARPDL(12,3,26, 8) Model

The forecast between 2010 and 2012 of the ARPDL(12,3,26,8) model is shown in Figure 5.30 and Figure 5.31.

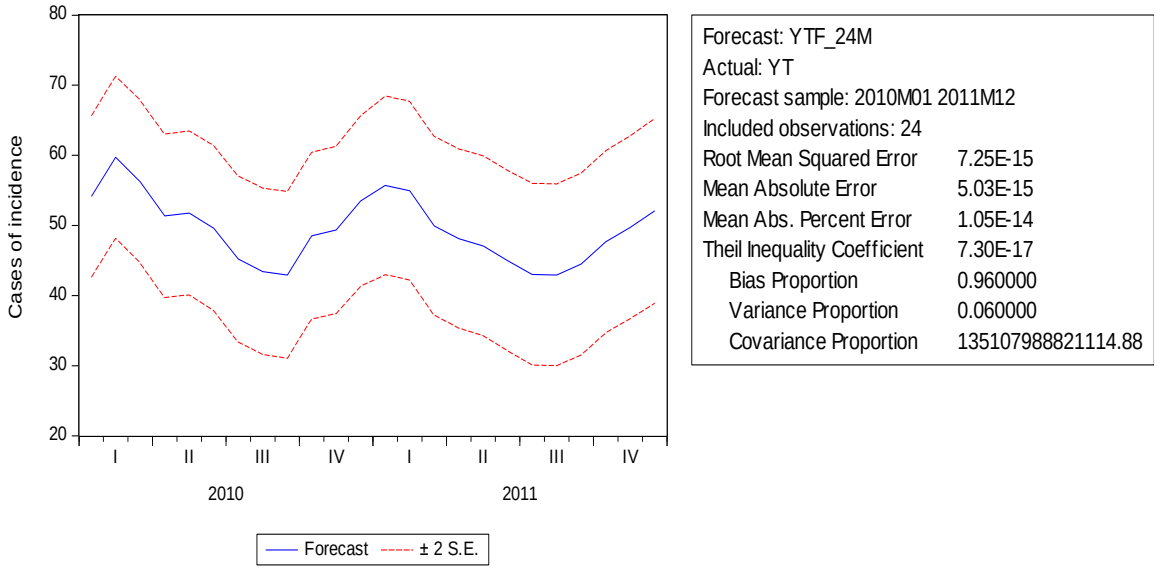


Figure 5.30: Forecast of the best ARPDL(12,3,26,8) model of lung cancer cases per month from 2010 to 2012.

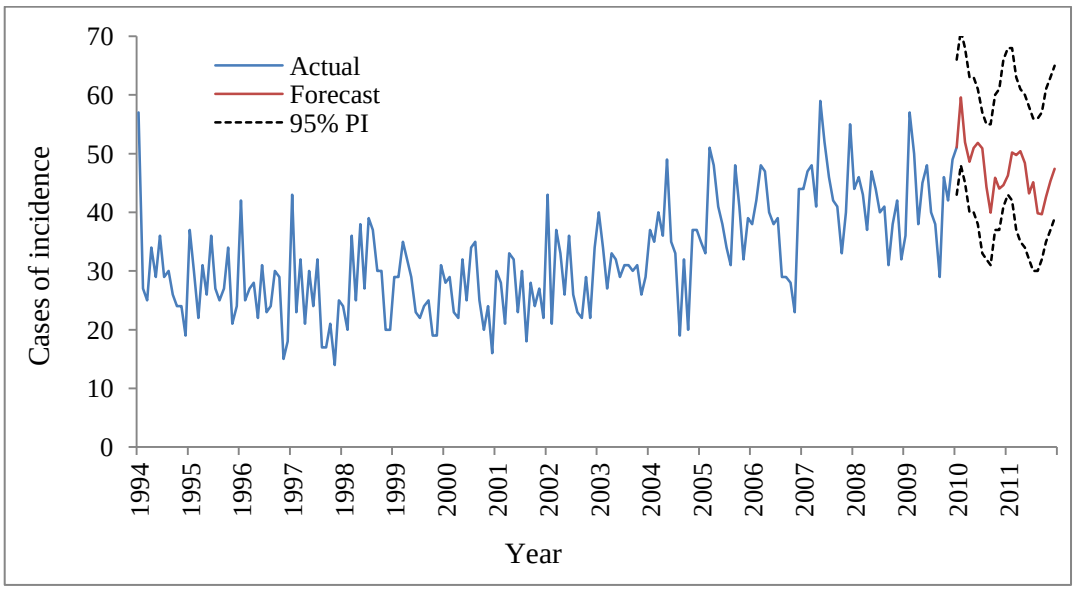


Figure 5.31: Actual and fitted ARPDL(12,3,26,8) model with 24 months ahead forecast of lung cancer cases per month from 1994 to 2012.

5.17. Discussion of Results

From Table 5.22, the best estimated dynamic model is ARPDL(12,3,26,8). This model has no autocorrelation, and the highest adjusted $\bar{R}^2$ and minimum forecast error among the dynamic ARPDL models fitted. The results of the short and long run effects are shown in Table 5.23. We now compare this model with the best SARIMA model determined in Chapter 4: this was the SARIMA(2,1,1)x$(0,1,1)_{12}$ model. Figure 5.32 shows the forecasts

generated by these two models. They both capture the seasonality trends reasonably well. However, the SARIMA model is preferred because it has a fewer parameters to estimate and only requires the past data on cases to define the forecast.

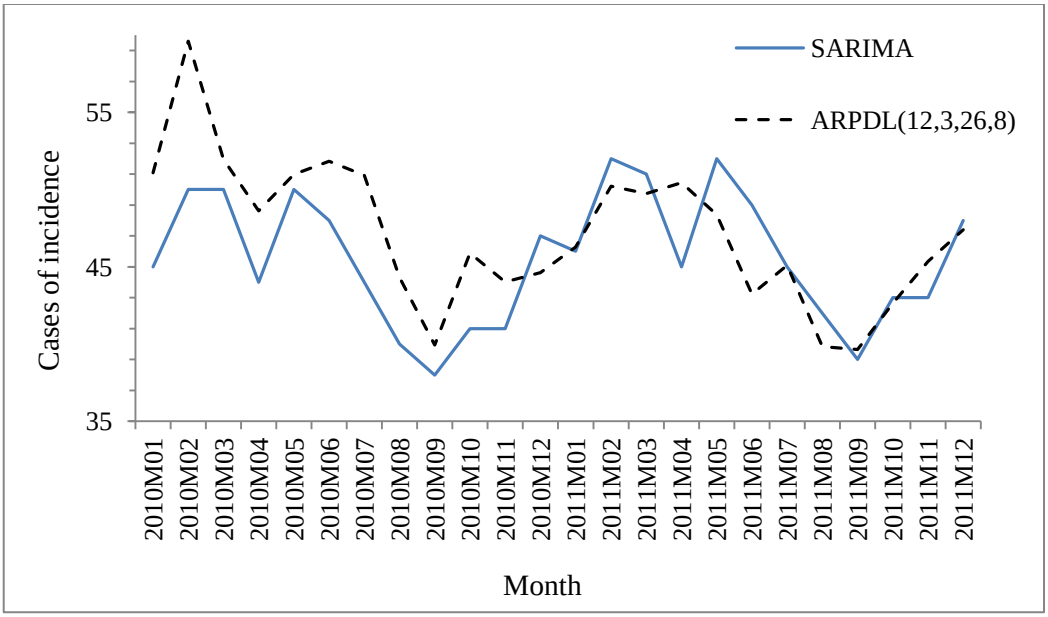


Figure 5.32: 24-step ahead forecast of lung cancer cases per month from 2010 to 2012 of best-fit SARIMA(2,1,1)x(0,1,1)$_{12}$ and ARPDL(12,3,26,8) models.

Table 5.22: Summary of Models I & II results.

Model	$\bar{R}^2$	DW	σ^2	One step ahead and 95%CI	Forecast Error
Model I:					
OLS (Best lag length = 26)	0.47	1.54	6447	53.8 (71.8, 35.7)	18.1
PDL(26,8)	0.48	1.70	7140	45.5 (59.1, 32.0)	13.5
ARPDL(12,5,26,8)	0.55	1.94	5917	50.0 (62.6, 37.5)	12.5
Model II:					
OLS (Best lag length = 26)	0.49	1.50	5022	60.5 (79.4, 41.5)	19.0
PDL(26,8)	0.50	1.83	6460	47.7 (61.0, 34.5)	13.2
ARPDL(12,3,26,8)**	0.62	2.02	4733	54.1 (65.6, 42.6)	11.5

ARPDL(12,3,26,8)** = Best dynamic model.

Table 5.23: Forecast cases of best ARPDL(12,3,26,8) model (2010-2011).

Month	Cases	Month	Cases
2010 Jan	54	2011 Jan	56
2010 Feb	60	2011 Feb	55
2010 Mar	56	2011 Mar	50
2010 Apr	51	2011Apr	48
2010 May	52	2011 May	47
2010 Jun	50	2011 Jun	45
2010 Jul	45	2011 Jul	43
2010 Aug	43	2011 Aug	43
2010 Sep	43	2011 Sep	44
2010 Oct	49	2011 Oct	48
2010 Nov	49	2011 Nov	50
2010 Dec	53	2011 Dec	52
Total	606		581

5.18. Summary

The data used are monthly incidence cases of lung cancer and smoking population for Saudi Arabia by gender from 1994-2009.

The empirical results suggest that lung cancer cases are strongly affected by smoking habits, and most of the cases are among males. However the value of the sum of t-ratios of the best model ARPDL(12,3,26,8) suggest that the smoking effect is greater for females than for males. The sum of the model coefficients also suggests that lung cancer cases decrease in males by 0.198 and increase in females by 1.893.

The one-step-ahead forecasts for each different model are:

1) Forecasting AR(1) model. The one-step-ahead forecast is 41 with 95% PI (26, 56). The mean square error is 69.5.
2) Forecasting linear regression model with lagged covariate. The one-step-ahead forecast is 44 with 95% PI (30, 57). The adjusted R-squared for the estimated relation is 45.3 and the mean square error is 45.42.
3) Forecasting linear regression model with lagged covariate and AR(1) errors. The one-step-ahead forecast is 43 with 95% PI (29, 56). The adjusted R-squared for the estimated relation is 36.4 and the mean square error is 69.5.
4) Forecasting distributed lagged variable model (DLM). The one-step-ahead forecast is 44 with 95% PI (30, 56). The adjusted R-squared for the estimated relation is 46.7 and the overall F-test value is 84.35 with p-value 0.00.

5) Forecasting of the best ARPDL(12,5,26,8) model of the total cases of lung cancer on total smoking population. The one-step-ahead forecast is 50 with 95% PI (37, 62). The adjusted R-squared for the estimated relation is 55.4 and the overall F-test value is 14.68 with p-value 0.00.
6) Forecasting of the best ARPDL(12,3,26,8) model of the total cases of lung cancer on smoking population separately for males and females. The one-step-ahead forecast is 54 with 95% PI (42, 65). The adjusted R-squared for the estimated relation is 62.6 and the overall F-test value is 13.55 with p-value 0.00.

The overall best one-step-ahead forecast was the total cases of lung cancer on smoking population separately for males and females ARPDL(12,3,26,8) model. This is confirmed by the value of adjusted R-squared as well as the significance of the F-statistic of the regression. Thus, the long run effect suggests that there will be on average 50 cases of lung cancer per month for the next 24 months. The estimated yearly lung cancer cases in 2010 and 2011 are 606 and 581 respectively. Subsequently, in winter (December - March), we have more incident cases being diagnosed (see Table 5.23).

Notice that our main aim of regressing the total cases of lung cancer on smoking population separately for males and females is that we want to identify the effect of changes in past values of smoking population separately for males and females on the current expected value of total cases of lung cancer. Particularly, we want to see where the effect of smoking is greater among males or females. In addition, we aim to minimize the error as much as possible since there are available data on smoking population for males and females separately in order to obtain reliable forecasts.

A new approach called Autoregressive Polynomial Distributed Lag (ARPDL) model was used to compare the errors associated with the model. In this approach, the procedure is the same as the PDL model except that we regressed Y_t on its polynomials as well. However, this procedure, which looks a little complex, was more flexible and parsimonious. It proved to be more robust on comparison with the PDL model approach, which is shown in the summary of results in Table 5.22. To the best of our knowledge, no study has been undertaken incorporating ARPDL approach to model and predict lung cancer incidence. This new procedure is outlined in section 5.14 with statistical software package Eviews8 commands.

Overall, ARPDL can be used when the number of observations available is limited and the number of significant lags is large. In this way, ARPDL models allow us to model

more complex lag structures of the independent and the dependent variables with different covariates but we need reasonable reasons for including them. ARPDL models are able to smooth the prediction and capture the seasonality trends. However, the forecast eventually becomes constant and does not predict a series with a seasonal pattern well.

Overall, when comparing the results obtained from SARIMA model to ARPDL model, we found that SARIMA model is preferred. Many advantages of SARIMA models were found and support the SARIMA model as a good way to forecast short-term time series. The SARIMA model has a fewer coefficients to estimate and only require the past data to define the forecast. Hence, SARIMA model can increase the forecast accuracy while keeping the number of parameters to a minimum.

One of the most widely used standard procedures for model evaluation in classification and regression is K-fold cross-validation (CV). However, when it comes to time series forecasting, because of the inherent serial correlation and potential non-stationarity of the data, its application is not straightforward and often omitted by practitioners in favour of an out-of-sample (OOS) evaluation. Hence we generated our forecasts accordingly using the seasonal ARIMA model.

It is important to mention that cross correlation methods in the time domain and impulse response functions in frequency domain which are generated through cross spectral analysis are other potential methods that can be used for modelling bivariate time series. Consideration of these approaches may lead to models that can be derived more efficiently than using lagged regression models with their many parameters. However, due to time constraints, we have not considered these approaches.

CHAPTER 6

AGE-PERIOD-COHORT MODELLING OF LUNG CANCER INCIDENCE

6.1. Introduction

Age-period-cohort (APC) models provide a useful method for modelling disease incidence and mortality rates of cancers (Rutherford et al., 2012). The effects of period and cohort are identified as proxies for events such as risk factors, which we cannot measure directly whereas the most important time-related variable that influences the risk of cancer is age (Bray and Moller, 2006). The age effect reflects the way of life, physiological, biological, behaviour factors, and lung risk factors for example. The period effect can highlight changes in the environmental factors that act around lung cancer onset including the effects of primary prevention and new medical care procedures. The cohort effect reflects the cumulative effects of exposure in generations (Meheni Khellaf, 2010).

APC models are known to suffer identification problems and that is due to the perfect relationship between the age, period, and cohort (Mason et al., 1973; Rutherford et al., 2010). This leads to a major challenge in analyzing APC models, a problem that has been widely addressed by statisticians, demographers and epidemiologists. The birth cohort can be calculated directly from the age at diagnosis and the date of diagnosis (cohort = period - age). If fitted directly in a generalized linear model (GLM) this leads to overparameterization and, consequently, incorrect parameter estimates because the APC model will not capture all the distinct effects of age, period, and birth cohort. It is therefore necessary to fit constraints to the model to extract identifiable answers for each of the parameters. This step is needed because each of the components of the model provides different insights into the trends of the disease over time.

New approaches have been developed for APC analysis to overcome the identification problem during the last 30 years. They are the conventional generalized linear CGLIM models and the intrinsic estimator IE (Yang, et al., 2004). In 2007, Carstensen developed new methodology for the identification problem. This author used age, period and cohort as continuous variables using spline functions. This author implemented this method for age-period-cohort models in R statistical software. In 2010, Rutherford et al developed a new command called `apcfit` that uses the spline functions, which was tested in STATA statistical software package. The identifiability problem is overcome by forcing constraints on either the period or the cohort effects.

Splines are a collection of polynomials that are joined at a pre-defined number of points known as knots. The first and last of these points are often referred to as the boundary knots. A spline is constrained in order to produce a smooth overall curve. It is worth noting that the number of knots determines the flexibility of the spline functions, which means that the number and location of knots can affect the fit. The function that is fitted is forced or restricted to have cubic curves between knots with continuous second derivatives at each knot and linear behaviour beyond the end knots. According to Sasieni (2012), because the splines are forced to be linear beyond the end knots, a natural cubic spline with no internal knots is simply a straight line (linear function).

Restricted cubic splines refer to restricted splines that use cubic polynomials between the knots and they have largely been used in other regression analyses according to Rutherford et al. (2012). In addition, cubic polynomials offer sufficient flexibility to capture the shape of most data, if appropriate knots are chosen.

In this chapter, we follow the procedure proposed by Rutherford et al. (2010). We outline the log-linear Poisson model in section 6.2 and present the APC modelling in section 6.3. In this analysis we present the APC basic model and include the covariates of gender, race, price of imported tobacco, consumption of tobacco per 1000 tons, smoking prevalence by gender, and five regions of Saudi Arabia. We present the overall best APC model with covariates in section 6.5.3. Prediction using restricted cubic (natural) splines and their graphs are presented in section 6.6.3. Finally, we discuss and give an overall summary of the chapter in section 6.7 and 6.8 respectively.

6.2. Log-linear Poisson Model

In GLMs, the dependent variable follows a distribution from the exponential family, which includes the normal, Poisson, binomial, exponential and gamma distributions (Montgomery et al., 2006, p. 160, 427). A GLM is a generalisation of the classical linear models (McCullagh and Nelder, 1983).

Taking into account that count data, like the lung cancer case data, are always non-negative, they are therefore naturally modelled on the log-scale. The choice of distribution to fit to the dependent variable is important. For count data not in the form of proportions, the Poisson distribution may be appropriate (McCullagh and Nelder, 1983, p. 127). For the Poisson distribution the variance is equal to the mean; Byers et al.,(2003) suggest that if the variance is much larger than the mean, a negative binomial distribution may be better suited to the data.

In our analysis, we use the method of model building as proposed by Clayton & Schifflers (1987). Rates are nonnegative and therefore are naturally modelled on the log-scale. The majority of the models fall into the class of generalized linear models and the assumptions often made are:

1. Assume that the count in each cell of the Lexis diagram y_{ij} is presented by Poisson (μ_{ijk}) with the expected rate $\lambda_{ijk} = \mu_{ijk}/N_{ij}$, where N_{ij} is the corresponding person–years at risk, i=1,...,m and j=1,..., n.
2. The person-years at risk (N_{ij}) is a fixed known value.
3. The random variables, y_{ij}, are jointly independent.
4. The expected rate is a logarithmic linear function as follows:

$$\ln(\lambda_{ijk}) = ln\left(\frac{\mu_{ijk}}{N_{ij}}\right) = \mu + \alpha_i + \beta_j + \gamma_k \qquad 6.1$$

where α_i represents the effect of age group i, β_j the effect of time period j, and γ_k the effect of birth cohort k. μ represents the mean effect or a constant corresponding to the log-rate for the reference levels (when i, j or k=*0*).

Typically, $ln\ (N_{ij})$ is treated as an offset when fitting a log-linear Poisson regression model with $\hat{\lambda}_{ijk} = y_{ij}/N_{ij}$ as the response. The persons at risk, N_{ij}, is not, strictly speaking, a fixed quantity. It is an estimation of a population collected from census registries each year from birth and death records, estimated immigration and emigration rates without any random variation in, N_{ij}, being many times larger than y_{ij}. Table F14 in Appendix F shows the person-years-at-risk (N_{ij}) in thousands of lung cancer cases among the entire population, aged 0-75+, in Saudi Arabia between 1994 and 2009.

Parameters can be estimated by means of a maximum likelihood procedure using statistical packages able to perform generalised linear modelling. Models are evaluated by their deviances from the null model, and then compared differences in deviance for model's best fit.

6.3. APC Modelling

Classically, APC models fit the effects of age, period, and cohort as factors. Due to a direct relationship between the terms, $cohort = period - age$, the components of this method cannot be uniquely determined. The models therefore need to be constrained in some way to ensure that the three functions showing age, period, and cohort effects can be extracted. Carstensen (2007), for example, demonstrated how this method could be achieved. The

method proposed by Carstensen uses restricted cubic (natural) splines for the age, period, and cohort terms within a GLM framework with a Poisson family error structure, a log link function, and an offset of log (person risk-time) to overcome the identification problem. According to Carstensen (2007), the APC model is to give an overview of the magnitude of the rates, the variation by age, and time trends in the rates.

However, in a slightly different method proposed by Rutherford et al (2010), transformations are made to the spline basis vectors for the period and cohort effects using matrix transformations. After successful transformation, a GLM is fitted within Stata using the adjusted spline basis vectors. Using this GLM as a foundation, it is possible to extend the analysis to include covariates. The data required to do this have observations for each unique age–period combination for every level of the covariate of interest. This allows us to adjust for the effect of the covariate by including the term in the GLM. It is also possible to include interaction terms between the covariate and age, period, and cohort.

Variation in lung cancer incidence could be explained by changes in smoking habits and other environmental risk factors such as air pollution, temperature, and price of tobacco that affect incidence rates and changes in risk factors that are present in early life. The models would identify and measure the effects of the age, period, and cohort on the disease incidence from 1994 to 2009.

We used the same approach as used by Rutherford et al. (2010) for the Lexis diagram to display the data. It summarizes a population's disease status over a calendar time against age. A Lexis diagram is usually split into five-year intervals for period and age. However it has been recommended by Rutherford et al. (2010) that yearly intervals should be used. Our data uses five-year age groups and cohorts but a one-year period interval. The data have been appropriately prepared in this way, each observation consisting of these explanatory variables: number of cases, population at risk, mean age, period and cohort. The command `poprisktime` was used to calculate the population risk-time from the population data using formula suggested by Sverdrup (1967) as in Carstensen 2007.

The incidence rates of lung cancer were calculated using the population of Saudi Arabia according to the statistical national census of 1994 to 2009 for all regions. It was decided to restrict the age range in 16 age classes of five years (0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75+), in 16 periods of one year (1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009), and in 91 cohorts of 5 years. To highlight the possibility of including covariates in the analysis, the gender and ethnicity of patients was included when

collapsing the dataset into unique records of age, period, and cohort. The incidence rate was measured as a function of age, period and cohort.

6.4. STATA Commands for Fitting APC Models

Two Stata commands `apc_ie` and `apc_cglim` are known to apply constraints to overcome the identifiability problem for the APC models. The `apc_ie` command uses the intrinsic estimator, which employs a principal components regression to arrive at the constrained estimates for the age, period, and cohort effects. The `apc_cglim` command on the other hand, uses a single equality constraint. The age, period, and cohort terms are fitted as factors, and a constraint that sets two of the categories from different components equal to one another is applied to overcome the lack of identifiability issue. The two approaches are described in detail and compared by Yang et al. (2004). In 2008, Land gave a good overview of techniques available to carry out APC models. Another command `apcfit`, which differs from the two approaches, uses restricted cubic splines to model the three variables and produces estimates for the three effects (age, period, and cohort) that can then be combined to give the predicted rates. These estimates can also be interpreted individually and plotted to show incidence and mortality trends over the different time scales. The advantage of `apcfit` is the potential for further modelling to investigate the effect of covariates (Rutherford et al., 2010).

6.5. Data Analysis and Results

6.5.1. The Basic Model

Having set up the data into the correct form, the `apcfit` command can now be applied. The `apcfit` command saves the adjusted spline basis as `_spA*` for the age variable, `_spP*` for the period variable, and `_spC*` for the cohort variable (see Table 6.1), which allows other models to be fit using the `glm` command (providing that the appropriate family, link, and offset are used). As a result, providing that the dataset was appropriately split for any given covariate, further models can be fit that can account for interactions.

Figure 6.1 shows the fitted incidence of lung cancer data for males and females combined. The default for `apcfit` is to make the reference point at the median value (with respect to the number of cases) for the period and cohort variables, respectively.

Table 6.1: The APC model of total lung cancer cases from 1994-2009.

Z	Coefficient	Standard Error	P-value	95% Confidence Interval	
				Lower	Upper
_spA1_intercept	-10.655	0.113	0.000	-10.875	-10.434
_spA2	2.376	0.132	0.000	2.118	2.634
_spA3	0.022	0.096	0.820	-0.167	0.210
_spA4	0.359	0.063	0.000	0.235	0.483
_spA5	0.076	0.029	0.010	0.018	0.133
_spA6	0.103	0.011	0.000	0.081	0.125
_spP1	-0.165	0.011	0.000	-0.186	-0.145
_spP2	-0.001	0.010	0.896	-0.022	0.019
_spP3	0.061	0.011	0.000	0.040	0.082
_spP4	0.030	0.010	0.003	0.010	0.050
_spC1_drift	-0.014	0.002	0.000	-0.019	-0.009
_spC2	0.005	0.086	0.958	-0.163	0.172
_spC3	-0.092	0.051	0.069	-0.192	0.007
_spC4	-0.016	0.051	0.759	-0.116	0.085
_spC5	-0.043	0.045	0.342	-0.130	0.045
ln (Y)	1.000	(exposure)			
Deviance = 4965.320		AIC = 4.889	Log likelihood = -4495.405		

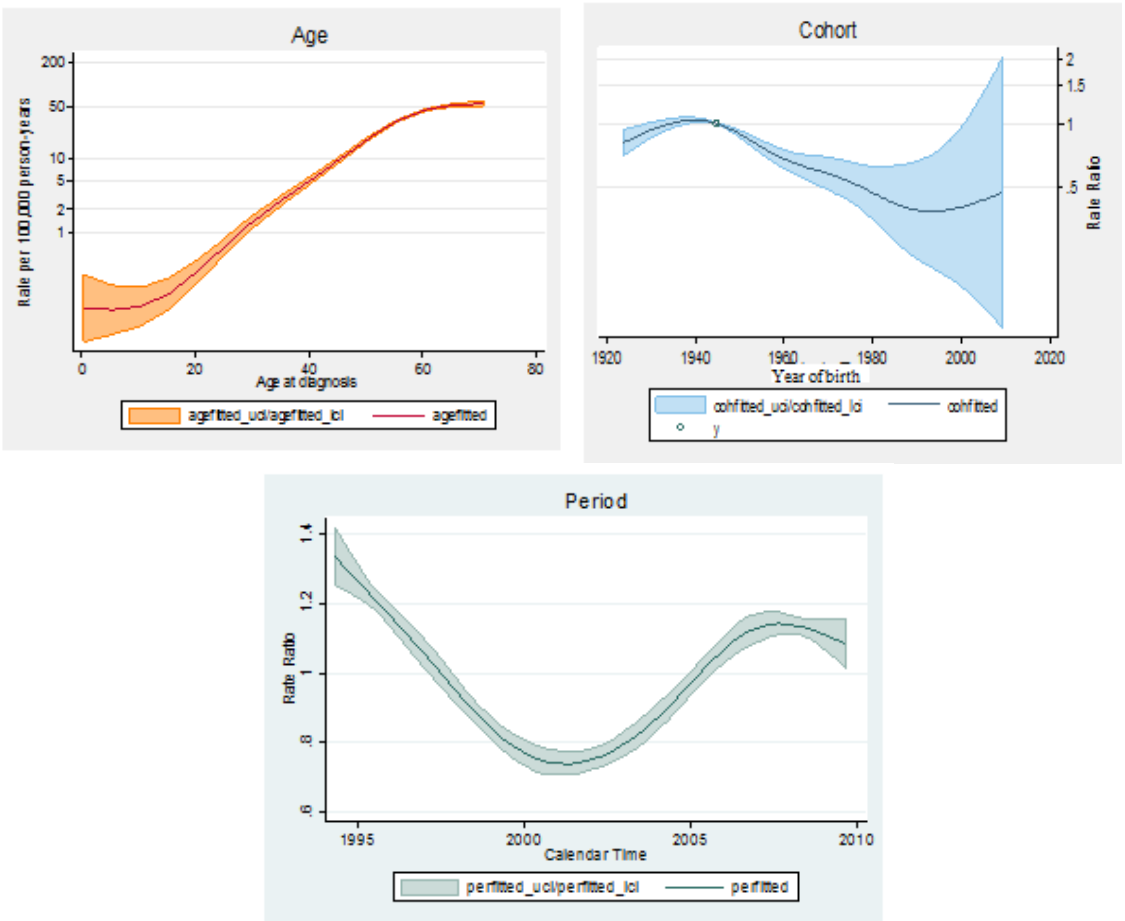


Figure 6.1: Age, cohort, and period effects of incidence rates for lung cancer data (degree of freedom=5) in Saudi Arabia. The respective regions surrounding the curves provides the 95% confidence bands. The circle indicates the reference point.

The incidence of cancer of the lung increases steadily from age 10 in the population up to the generation born before the second world war in 1939 and declines thereafter until the Gulf war in 1990. A subsequent increase followed in cohorts born after the Gulf war until 2010. However these figures are based on small numbers. Actually this complex cohort effect might be over fitting the splines. There may in fact be simply a decrease in incidence with cohort as lifestyles become healthier overall. The second increase in cohort incidence we see in the 1990s may be due to the establishment of Saudi Cancer Registry, influx of foreign nationals or immigration because of the Gulf war and the availability of newer diagnostic techniques using computers. The period of incident rates declined for about a decade to the early 2000s and then rose up to 2007, and thereafter observed a subsequent decrease.

The degrees of freedom were set to five (default) for each of the spline bases for the age, period, and cohort. It is interesting to alter the degrees of freedom for any one of the variables, particularly the cohort variable, although this might lead to over fitting if the number is increased too much. The decision on the number of degrees of freedom can be aided using the Akaike's information criterion (AIC) values. A lower AIC value suggests a better fitting model.

6.5.2. Computation of AIC and BIC Computed in Stata

Stata commands `glm`, `binreg`, and `ml` use the following formulae to compute the values of AIC and BIC:

$$\mathrm{AIC} = (-2\mathrm{lnL} + 2\mathrm{k})/\mathrm{N}$$

$$\mathrm{BIC} = \mathrm{D2} - (\mathrm{N} - \mathrm{k})\ln(\mathrm{N})$$

where `lnL` and `D2` are the overall likelihood and the overall deviance, reported by `glm`, `k` is the number of parameters of the model, and `N-k` is the degrees of freedom associated with the deviance `D2`. These formulae are from Akaike (1973) and Raftery (1995), respectively.

On the other hand, Stata commands `estat ic`, `estimates table`, and `stats (aic bic)` use different definitions of these criteria on the basis of Akaike (1974) and Schwarz (1978):

$$\mathrm{AIC} = -2\mathrm{lnL} + 2\mathrm{k}$$

$$\mathrm{BIC} = -2\mathrm{lnL} + \mathrm{kln}\,(\mathrm{N})$$

They thus report different AIC and BIC values.

6.5.3. Inclusion of Covariates

We first estimated generalized log linear models in Equation (6.1) for the total lung cancer incidence data in Saudi Arabia. Using the GLM command, we can then add terms to the model to take into account the effects of other covariates and hence the best model. In this section, various combinations of covariates such as gender, race, consumption of tobacco per 1000 tons, price of imported tobacco, smoking prevalence for males and females, and five regions (north, south, east, west, central) of Saudi Arabia were added to assess the performance of the model.

The essence of including various combinations of covariates was to demonstrate the preferred model for the lung cancer incidence in KSA.

We started by estimating the reduced models with the covariates, and then the full three-way APC model. The marginal or gross effects for each of the components of age (A) and period (P) with their model selection statistics are given in Table 6.2 and Table 6.3 respectively. The two-way models of age & period (AP), age & cohort (AC), and period & cohort (PC) with their model selection statistics are given in Table 6.4 through to Table 6.6. Finally, we estimated the full three-way APC model.

In order to fit an identified APC model, we considered the best model fit for each of the six models with their covariates using the model selection statistics. The best-fit models are indicated in the tables by double stars (**).

Table 6.2: Covariates with age (A) model.

Model	Deviance	Pearson	DF	Log Likelihood	AIC
A	5289.62	6091.68	1839	-4657.55	5.055
A+GENDER	3225.11	4796.36	1838	-3625.30	3.937
A+ GENDER +RACE	3206.72	4613.17	1837	-3616.10	3.928
A+ GENDER +RACE+CONSUMPTION	3133.71	4497.26	1836	-3579.60	3.890
A+ GENDER +RACE+CONSUMPTION+PRICE	3101.95	4511.14	1835	-3563.72	3.873
A+ GENDER +RACE+CONSUMPTION+PRICE+SMOKING	2973.66	4339.68	1834	-3499.57	3.805
A+ GENDER +RACE+CONSUMPTION+PRICE+SMOKING +(FIVE REGIONS)*	2948.29	4407.62	1829	-3486.89	3.797
A+ GENDER +RACE+CONSUMPTION+SMOKING+FIVE REGIONS	2948.48	4406.95	1830	-3486.99	3.796
A**+RACE+CONSUMPTION+SMOKING+FIVE REGIONS	2950.10	4376.92	1831	-3487.80	3.795
A+CONSUMPTION+SMOKING+FIVE REGIONS	2961.16	4526.29	1832	-3493.33	3.800

(FIVE REGIONS)* = NORTHERN, SOUTHERN, EASTERN, WESTERN, AND CENTRAL
A**= BEST MODEL FOR AGE

Table 6.3: Covariates with period (P) model.

Model	Deviance	Pearson	DF	Log Likelihood	AIC
P	611643747.4	305891505.6	1841	-305823886	331516.4
P**+ GENDER	268083230.5	134136604.2	1840	-134043628	145304.8

P**= BEST MODEL FOR P

Table 6.4: Covariates with age-period (AP) model.

Model	Deviance	Pearson	DF	Log Likelihood	AIC
AP	5033.26	5752.25	1835	-4529.38	4.920
AP+ GENDER	2955.44	4629.97	1834	-3490.46	3.795
AP+ GENDER +RACE	2943.52	4466.11	1833	-3484.50	3.790
AP+ GENDER +RACE+CONSUMPTION	2920.08	4420.79	1832	-3472.78	3.778
AP+ GENDER +RACE+CONSUMPTION+PRICE	2919.52	4423.67	1831	-3472.51	3.779
AP+ GENDER +RACE+CONSUMPTION+PRICE+ SMOKING	2916.58	4388.80	1830	-3471.03	3.778
AP+ GENDER +RACE+CONSUMPTION+SMOKING	2916.58	4388.16	1831	-3471.04	3.777
AP+ GENDER +RACE+CONSUMPTION+SMOKING+ FIVE REGIONS	2902.36	4427.85	1826	-3463.92	3.774
AP**+ GENDER +RACE+CONSUMPTION+ FIVE REGIONS	2899.13	4431.94	1827	-3462.31	3.772

AP**= BEST MODEL FOR AP

Table 6.5: Covariates with age-cohort (AC) model.

Model	Deviance	Pearson	DF	Log Likelihood	AIC
AC	5253.13	6024.78	1834	-4639.31	5.040
AC+ GENDER	3185.18	4754.19	1833	-3605.33	3.921
AC+ GENDER +RACE	3166.80	4571.88	1832	-3596.14	3.912
AC+ GENDER +RACE+CONSUMPTION	3071.11	4503.47	1831	-3548.30	3.861
AC+ GENDER +RACE+CONSUMPTION+PRICE	3066.05	4512.61	1830	-3545.77	3.859
AC+ GENDER +RACE+CONSUMPTION+PRICE+ SMOKING	2922.82	4335.89	1829	-3474.15	3.783
AC+ GENDER +RACE+CONSUMPTION+SMOKING	2923.36	4332.18	1830	-3474.42	3.782
AC+ GENDER +RACE+CONSUMPTION+SMOKING+ FIVE REGIONS	2860.83	4485.99	1825	-3443.16	3.754
AC+ GENDER +RACE+CONSUMPTION+ FIVE REGIONS	2861.46	4460.69	1826	-3443.47	3.753
AC**+ GENDER +RACE+FIVE REGIONS	2861.93	4458.62	1827	-3443.71	3.752
AC+ GENDER +RACE+PRICE+CONSUMPTION+ FIVE REGIONS	2861.44	4461.48	1825	-3443.47	3.754

AC**= BEST MODEL FOR AC

Table 6.6: Covariates with period-cohort (PC) model.

Model	Deviance	Pearson	Obs	DF	Log Likelihood	AIC
PC 4700.717 Warning: convergence not achieved	8668779.493	1.8669e+13	1845	1836	-4336402.491	

Table 6.7: Covariates with age-period-cohort (APC) model.

Model	Deviance	Pearson	DF	Log likelihood	AIC
APC	4965.32	5674.70	1830	-4495.40	4.889
APC+ GENDER	2885.20	4599.72	1829	-3455.34	3.762
APC+ GENDER +RACE	2873.75	4438.14	1828	-3449.62	3.757
APC+ GENDER +RACE+CONSUMPTION	2866.83	4434.29	1827	-3446.16	3.755
APC+ GENDER +RACE+CONSUMPTION+PRICE	2866.83	4434.04	1826	-3446.16	3.756
APC+ GENDER +RACE+CONSUMPTION+SMOKING	2865.41	4409.41	1826	-3445.45	3.755
APC+ GENDER +RACE+CONSUMPTION+FIVE REGIONS	2848.15	4446.25	1822	-3436.82	3.750
APC**+ GENDER +RACE+FIVE REGIONS (1)	2849.07	4444.55	1823	-3437.28	3.749
APC+ GENDER +RACE+SMOKING	2870.26	4400.82	1827	-3447.87	3.757
APC+ GENDER +FIVE REGIONS	2860.42	4608.90	1824	-3442.95	3.754
APC*+ GENDER +RACE+PRICE+CONSUMPTION+ SMOKING+FIVE REGION	2846.05	4478.39	1820	-3435.77	3.751

APC**= BEST MODEL FOR APC

Table 6.8: The best five models.

Model	Deviance	Pearson	DF	Log likelihood	AIC
A+RACE+CONSUMPTION+SMOKING+FIVE REGIONS	2950.10	4376.92	1831	-3487.80	3.795
AP+ GENDER +RACE+CONSUMPTION+FIVE REGIONS	2899.13	4431.94	1827	-3462.31	3.772
AC+ GENDER +RACE+FIVE REGIONS	2861.93	4458.62	1827	-3443.71	3.752
APC+ GENDER +RACE+PRICE+CONSUMPTION+SMOKING +FIVE REGIONS	2846.05	4478.39	1820	-3435.77	3.751
APC+ GENDER +RACE+FIVE REGIONS ***	2849.10	4444.55	1823	-3437.28	3.749

*** = Overall Best Model

From Table 6.2, it is clear that race, consumption, smoking and the five regions influence the age effects most. From Table 6.3, gender is the only covariate that influences the period effects. By contrast to Table 6.2, the age-period (AP) models in Table 6.4 provide better results. It is therefore clear that gender, race, consumption, and the five regions influence the age-period effects most. From Table 6.5, we can also see that gender, race, and the five regions influence the age-cohort effects most. There was no convergence in any of the covariates or combination of covariates with the period-cohort (PC) model. Finally, when we used the full age-period-cohort model, we realized that gender, race, and the five regions best influenced the age-period-cohort effects.

The model selection statistics reported in Table 6.8 for each of these five best models selected from both reduced and full three-way APC models show that the three full APC models with the covariates of gender, race, and five regions fit the data significantly better than other four models. This is indicated in Table 6.8 as triple star (***). The results from the best five models are presented in Table 6.9.

Therefore, it can be concluded that none of the three components of the APC model should be eliminated from the model specification and selection. Hence, we present Table 6.10 that depicts the coefficient estimates and model fit statistics of the overall best model.

Table 6.9: The best five models with different covariates.

	APC+GENDER +RACE+FIVE REGION			APC+GENDER +RACE+PRICE+ CONSUMPTION+ SMOKING+ FIVEREGIONS			A+RACE+ CONSUMPTION +SMOKING+FIVE REGIONS			AP+GENDER +RACE+ CONSUMPTION + FIVE REGIONS			AC+GENDER +RACE+ FIVE REGIONS		
Z	IRR	Std.	P>|z|	IRR	Std.	P>|z|	IRR	Std.	P>|z|	IRR	Std.	P>|z|	IRR	Std.	P>|z|
_spA1_intct	0.00	0.00	0.000	0.00	0.00	0.000	0.00	0.00	0.000	0.00	0.00	0.000	0.00	0.00	0.000
_spA2	5.73	1.56	0.000	8.21	3.94	0.000	15.02	0.67	0.000	15.04	0.67	0.000	3.18	0.51	0.000
_spA3	0.99	0.09	0.880	0.99	0.09	0.878	0.95	0.04	0.219	0.95	0.04	0.221	0.99	0.10	0.913
_spA4	1.42	0.09	0.000	1.42	0.09	0.000	1.49	0.07	0.000	1.49	0.07	0.000	1.42	0.09	0.000
_spA5	1.09	0.03	0.004	1.09	0.03	0.004	1.11	0.02	0.000	1.11	0.02	0.000	1.09	0.03	0.004
_spA6	1.12	0.01	0.000	1.12	0.01	0.000	1.17	0.01	0.000	1.17	0.01	0.000	1.13	0.01	0.000
_spP1	0.93	0.02	0.004	0.94	0.03	0.072	-	-	-	0.96	0.03	0.258	-	-	-
_spP2	0.97	0.02	0.077	0.94	0.03	0.058	-	-	-	0.92	0.02	0.000	-	-	-
_spP3	1.03	0.02	0.123	1.05	0.03	0.045	-	-	-	1.06	0.02	0.004	-	-	-
_spP4	1.03	0.02	0.224	1.03	0.03	0.292	-	-	-	1.02	0.03	0.452	-	-	-
_spC1_ldrft	0.96	0.01	0.000	0.97	0.02	0.198	-	-	-	-	-	-	0.93	0.00	0.000
_spC2	0.98	0.08	0.857	0.99	0.08	0.860	-	-	-	-	-	-	0.98	0.08	0.815
_spC3	0.91	0.05	0.076	0.91	0.05	0.078	-	-	-	-	-	-	0.92	0.05	0.084
_spC4	0.99	0.05	0.869	0.99	0.05	0.876	-	-	-	-	-	-	0.99	0.05	0.879
_spC5	0.96	0.04	0.362	0.96	0.04	0.361	-	-	-	-	-	-	0.96	0.04	0.381
gender	2.79	0.07	0.000	4.46	1.73	0.000	-	-	-	2.79	0.07	0.000	2.79	0.07	0.000
race	0.92	0.02	0.001	0.92	0.02	0.001	0.92	0.02	0.001	0.92	0.02	0.001	0.92	0.02	0.001
consumption	-	-	-	1.00	0.01	0.761	0.98	0.00	0.000	0.98	0.00	0.000	-	-	-
price	-	-	-	1.00	0.00	0.335	1.07	0.00	0.000	-	-	-	-	-	-
smoking	-	-	-	0.97	0.03	0.226	1.07	0.00	0.000	-	-	-	-	-	-
Northern	1.00	0.00	0.277	1.00	0.00	0.297	1.00	0.00	0.899	1.00	0.00	0.217	1.00	0.00	0.780
Southern	1.01	0.00	0.044	1.01	0.00	0.056	1.00	0.00	0.253	1.00	0.00	0.184	1.01	0.00	0.040
Western	1.00	0.00	0.030	1.00	0.00	0.038	1.00	0.00	0.000	1.00	0.00	0.065	1.00	0.00	0.000
Central	1.00	0.00	0.471	1.00	0.00	0.120	1.00	0.00	0.662	1.00	0.00	0.378	1.00	0.00	0.000
Eastern	1.00	0.00	0.038	1.01	0.00	0.043	1.00	0.00	0.037	1.00	0.00	0.187	1.00	0.00	0.000
Deviance	2849.066			2846.049			2950.104			2899.135			2861.928		
Person	4444.545			4478.391			4376.924			4431.944			4458.616		
Log likelihood	-3437.277			-3435.769			-3487.796			-3462.312			-3443.708		
AIC	3.750			3.752			3.796			3.773			3.753		

Table 6.10: Overall best APC model.

Z	IRR	Standard Error	P-value	95% Confidence Interval	
				Lower	Upper
spA1 intercept	0.000	0.000	0.000	0.000	0.000
_spA2	5.727	1.556	0.000	3.362	9.755
_spA3	0.986	0.095	0.880	0.817	1.190
_spA4	1.421	0.090	0.000	1.256	1.608
_spA5	1.089	0.032	0.004	1.028	1.153
_spA6	1.125	0.013	0.000	1.100	1.150
_spP1	0.929	0.024	0.004	0.883	0.977
_spP2	0.973	0.015	0.077	0.944	1.003
_spP3	1.033	0.022	0.123	0.991	1.076
_spP4	1.029	0.024	0.224	0.983	1.077
_spC1_ldrift	0.957	0.011	0.000	0.936	0.978
_spC2	0.985	0.084	0.857	0.833	1.164
_spC3	0.914	0.046	0.076	0.828	1.009
_spC4	0.992	0.051	0.869	0.897	1.096
_spC5	0.960	0.043	0.362	0.880	1.048
gender	2.793	0.070	0.000	2.659	2.934
race	0.919	0.023	0.001	0.876	0.965
Northern	0.996	0.004	0.277	0.989	1.003
Southern	1.006	0.003	0.044	1.000	1.012
Western	1.003	0.001	0.030	1.000	1.005
Central	1.001	0.001	0.471	0.999	1.003
Eastern	1.004	0.002	0.038	1.000	1.009
ln(Y)	1.000	(exposure)			
Deviance = 2849.06		AIC = 3.750	Log likelihood = -3437.27		

The simplest method for the inclusion of the gender term, for example, as covariate into the GLM is to assume a proportional effect for gender. The covariate for gender is coded as 0 for female and 1 for male. The `eform` option in Stata is used to report the covariate terms as an incidence rate ratio (IRR). In gender for example, we look at the effects of males relative to females. Similarly, the covariate for race is coded as 0 for non-Saudi and 1 for Saudi. Thus, we also look at the effects of non-Saudis relative to Saudis.

The output given above shows that, in KSA, males have about a 79% greater incidence of lung cancer than females across the entire dataset when adjusting for the other effects. The p-value for the gender term highlights that the effect for gender is significant at the 5% level and even at the 0.1% level. This measure of significance, however, assumes that the effect of gender is proportional over both time scales and date of birth. In addition, the p-values for the race, Southern region, Western region, and Eastern region terms show that the effect for these covariates is statistically significant.

6.6. Prediction Using Restricted Cubic (Natural) Splines

6.6.1. Introduction

Many methods have been proposed for making predictions from APC models. The technical aspects of forecasting the burden of cancer have been developed and refined over the past few decades. For more information on APC model projections see the following papers, Bray et al. (2001), Moller et al. (2003), Clements et al. (2005), Bray and Moller (2006), Carstensen (2007), Cleries et al. (2010), Rutherford et al. (2010), Lee et al. (2011), Mistry et al. (2011), Rutherford et al. (2012) and Sasieni (2012). Natural cubic splines were firstly used in APC models by Sasieni and Adams (1999, 2000) for drawing inference on the impact of cervical screening on cervical cancer rates. Quite apart from these methods, good overviews of techniques available to carry out APC model projections using natural cubic splines have been given by Rutherford et al. (2012) and Sasieni (2012). In 2012, Rutherford et al. and Sasieni summarised that multiplicative APC models tend to over-estimate future rates of a disease incidence or mortality and therefore linear projections need to be tempered or dampened when making long-term predictions. For that reason, they advocated the use of an APC with a power link function together with a linear combination of age, period and cohort terms.

Although the `apcfit` uses a canonical link, using predict after fitting a `glm` command does not give you correct fitted mean values. This is because predict after `glm` does not take care of the regularization of the background or smoothing the model except the default background only. Another reason is that `poisson` or `apcfit` is not fully flexible and does not facilitate the visualization of the functions of age, period and cohort effects. Hence, making projections or forecasting from such predicted or fitted values could be very misleading. Notably, `apcfit` is used for fitting APC models using natural cubic splines when not making projections. Hence, an associated command `apcspline` will make projections simpler from `apcfit`.

What makes `apcspline` more effective and powerful for making projections is that different link functions can be used on the rates. By using the `apcspline`, a trade-off or balance also exists between having the flexibility to capture the salient features of the cohort effect and having a parsimonious model.

In the `apcspline` command, constraints are imposed by centering the period effects and cohort effects at the mean year of cases and at the weighted mean year of birth respectively, whereby `apcfit` allows the user to specify the centering of each variable.

6.6.2 The APC Model Prediction

The `apcspline` command fits an APC model of the form

$$Z \sim \text{Poisson}$$

$$g\left(\frac{\mu}{\text{exposure}}\right) = f_A(\text{age}) + f_P(\text{period}) + f_C(\text{cohort}) + \beta_{\text{drift}} \qquad 6.2$$

where g is the link function, μ is the mean of the incidence rate and f_A, f_P, and f_C are the natural cubic splines. β_{drift} is the common drift parameter (Clayton & Schifflers, 1987).

When comparison between age only and age-period models indicates a highly significant period effect, and comparison between age only and age-cohort models indicates a highly significant cohort effect, we conclude that both age-period and age-cohort models fit the data very well. An explanation is that there is some temporal variation of rates which does not distinguish between period and cohort influences; that is, a variation over time which could be predicted either by the age-period model or by the age-cohort model. This interesting phenomenon of a data set described equally well by both models is known as drift (Clayton and Schifflers, 1987).

The main predictions illustrated here are based on both the power of 0.2 link function and on the spline functions of age, period and cohort. Other functions were used to study the sensitivity of the results to these constraints. We use the world standard population presented by Doll et al. (1966) which is the most frequently standard population used for the age standardization.

According to Sasieni (2012) the power 0.2 link function is used in our analysis of this thesis to reduce the growth in the predicted rates. It has been found that for moderate trends the difference between the logarithmic and the power 0.2 link in terms of fitted values to the observed numbers of events will be minimal, but the impact on long-term extrapolation could be considerable.

The data we use to illustrate the `apcspline` command contain the number of cases of lung cancer in Saudi Arabia in 5-year age bands for each year from 1994–2009 together with mid-year population estimates for 1994–2009 and population projections until 2020. The numbers of both lung cancer cases and population are separated by gender. For comparison, we fit both `apcspline model` and `apcfit model` and present the results in Table 6.11 and in Table 6.12 respectively.

Table 6.11: `apcspline` model for male lung cancer from 1994-2009.

Z	Coefficient	Standard Error	P-value	95% Confidence Interval	
				Lower	Upper
A	0.138	0.005	0.000	0.127	0.148
_IA1	-0.001	0.001	0.026	-0.002	0.000
_IA2	0.001	0.001	0.101	0.000	0.002
_IA3	0.000	0.000	0.245	-0.001	0.000
_IA4	0.000	0.000	0.146	0.000	0.001
_IA5	0.000	0.000	0.008	-0.001	0.000
_IA6	0.000	0.000	0.000	0.000	0.001
β_{drift}	-0.022	0.004	0.000	-0.030	-0.014
_IP1	0.010	0.008	0.225	-0.006	0.025
_IP2	-0.004	0.011	0.709	-0.026	0.018
_IP3	-0.008	0.011	0.477	-0.031	0.014
_IP4	0.001	0.010	0.906	-0.019	0.022
_IP5	0.010	0.007	0.169	-0.004	0.024
_IC1	0.000	0.000	0.838	0.000	0.000
_IC2	0.000	0.000	0.168	0.000	0.000
_IC3	0.000	0.000	0.046	0.000	0.000
_cons	-16.709	0.295	0.000	-17.287	-16.131
ln(population)	1.000	(exposure)			
Log likelihood = -601.14016 Predict fitapc (option n assumed; predicted number of events)					

Table 6.12: `apcfit` model for male lung cancer from 1994-2009.

Z	Coefficient	Standard Error	P-value	95% Confidence Interval	
				Lower	Upper
_spA1_intercept	-10.810	0.188	0.000	-11.178	-10.442
_spA2	2.396	0.230	0.000	1.946	2.846
_spA3	0.154	0.176	0.380	-0.190	0.499
_spA4	0.334	0.108	0.002	0.121	0.546
_spA5	0.089	0.046	0.051	0.000	0.179
_spA6	0.106	0.017	0.000	0.073	0.139
_spP1	-0.182	0.015	0.000	-0.212	-0.153
_spP2	0.038	0.014	0.009	0.009	0.066
_spP3	-0.006	0.015	0.670	-0.036	0.023
_spP4	0.085	0.015	0.000	0.055	0.115
_spC1_drift	-0.021	0.003	0.000	-0.028	-0.015
_spC2	-0.110	0.143	0.443	-0.391	0.171
_spC3	-0.226	0.084	0.007	-0.390	-0.062
_spC4	-0.111	0.083	0.182	-0.274	0.052
_spC5	-0.098	0.071	0.163	-0.237	0.040
ln (population)	1.000	(exposure)			
Log likelihood = -607.68 Predict fitapc (option mu assumed; predicted mean Z) (608 missing values generated)					

Note that the fitted values from `apcfit` are only available for the observations that were used in the model fitting, whereas predict after `apcspline` provides estimated mean numbers for all observations. However, the fitted values that are provided by both commands are extremely similar. The cohort effect estimated by `apcspline` command is not the same as that estimated by `apcfit`. This is due to the effect of the transformation used by each command.

Table 6.13: Comparison between `apcspline` and `apcfit` command.

Variable	Observation	Mean	Standard deviation	Min	Max
fit apcspline	256	18	18.76	0.14	60.92
fit apcfit	256	18	18.86	0.17	64.22

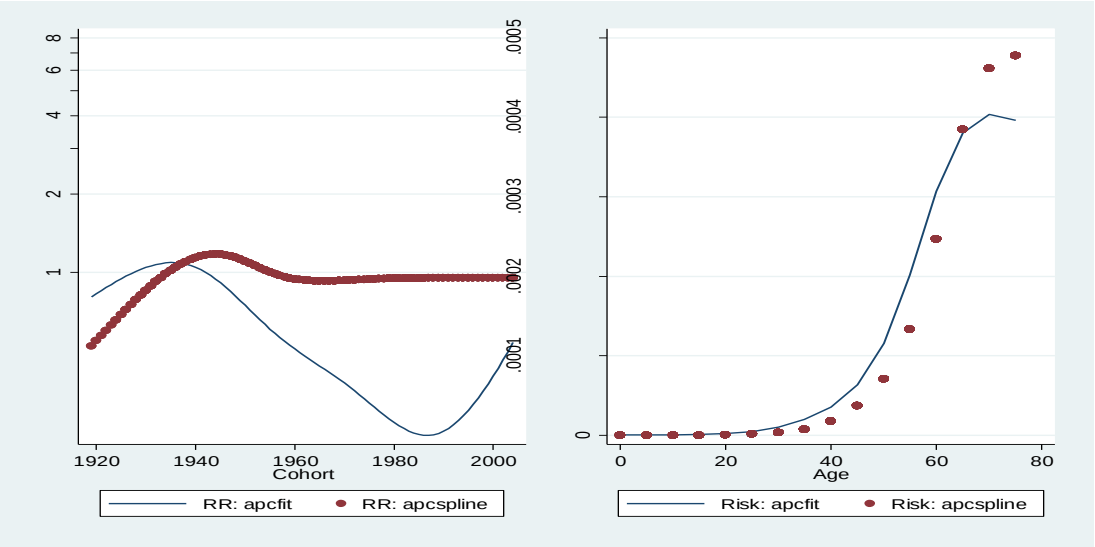


Figure 6.2: Comparison of the default output from `apcspline` with that from `apcfit`.

It can be seen that the estimated risks as a function of age are similar, but the cohort relative risks are quite different. The left-hand plot shows the estimated cohort effects, which are very different. In particular, in the `apcspline` model fit, the relative risk is always close to 1, whereas the `apcfit` gives an estimate that decreases rapidly to beyond 1 for those born between 1940 and 1980 and then increases. It should be noted that the constraints imposed by the two programs are different: one could remove the drift from the `apcfit` cohort effect, but its tail behaviour would still be quite different from the `apcspline` effect. The right-hand plot compares the age curve from both models. They

are seen to be similar. Thus, we prefer the apcspline to the apcfit for forecasting purposes because the apcspline command is flexible and captures the salient features of the cohort effect according to Sasieni (2012).

The `apcspline` command can also be used to generate the bases for the splines, which can then be combined with other covariates or multiplied to produce interactions within a `Poisson` or `glm` model.

6.6.3. Graphs: Spline Predictions

Figure 6.3 shows the observed age specific standardised incidence rates plotted for males and females separately from 1994-2009 with fitted rates, and predictions of rates from 2010-2020 derived from the APC model using the spline functions. For example, see Figure B1 in Appendix B1. From Figure 6.3, the cause of the bump in risk for males in 2007 is likely due to the history of high smoking prevalence among males in that period (see the smoking population in Figure 3.3). Note that we can use more than one model fit, as shown in Figure 6.4. We can also do cohort plots (see Figure 6.5).

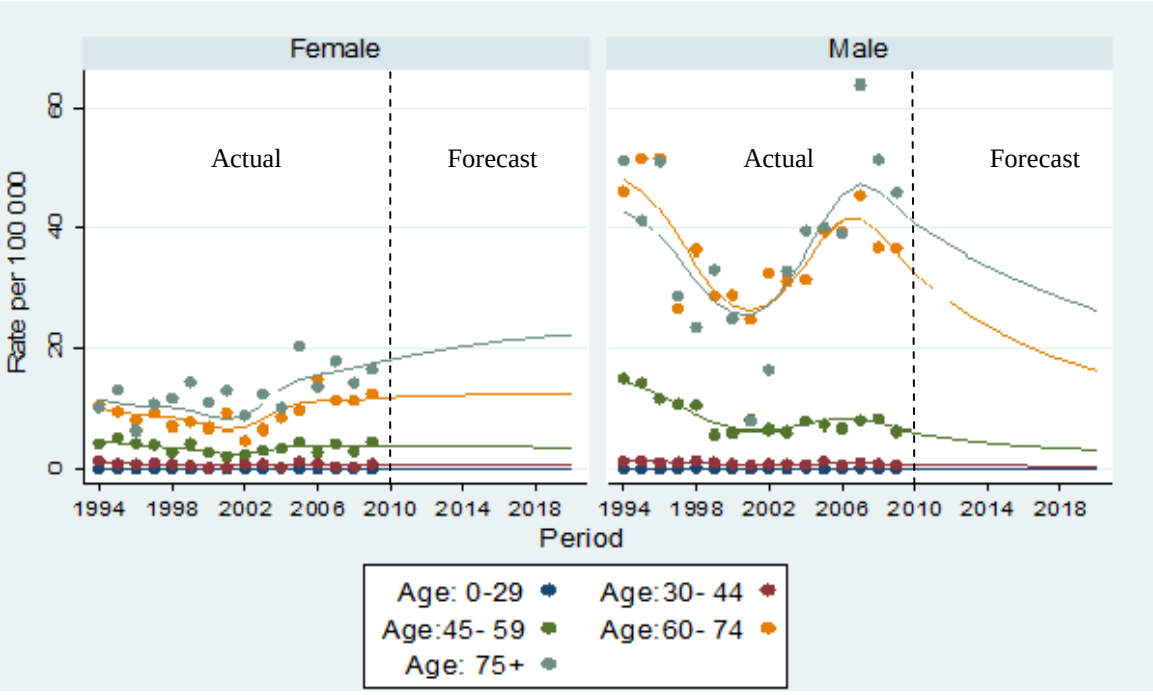


Figure 6.3: Actual (solid circles ••••) and fitted (solid curve) age-specific standardised rates of lung cancer incidence in KSA (per 100,000 person-year) from 1994 to 2009 with forecast rates from 2010 to 2020 for males and females separately with different age bands.

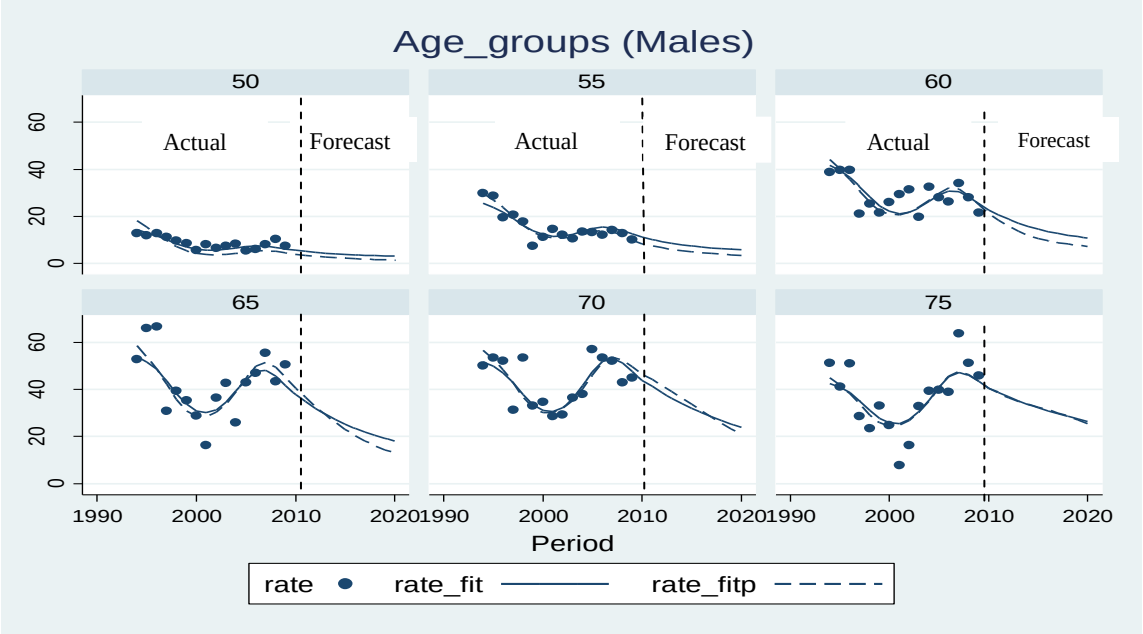


Figure 6.4: Actual (solid circles ••••) age-specific standardised rates of lung cancer incidence (per 100,000 person-year) with the fitted rate from 1994 to 2009 and the projected rate from 2010 to 2020 for males in KSA for age groups 50-75 years. Both the predictions based upon the logarithmic link (solid curve) and the predictions based on the power 0.2 link (dashed curve) are shown. They are almost identical.

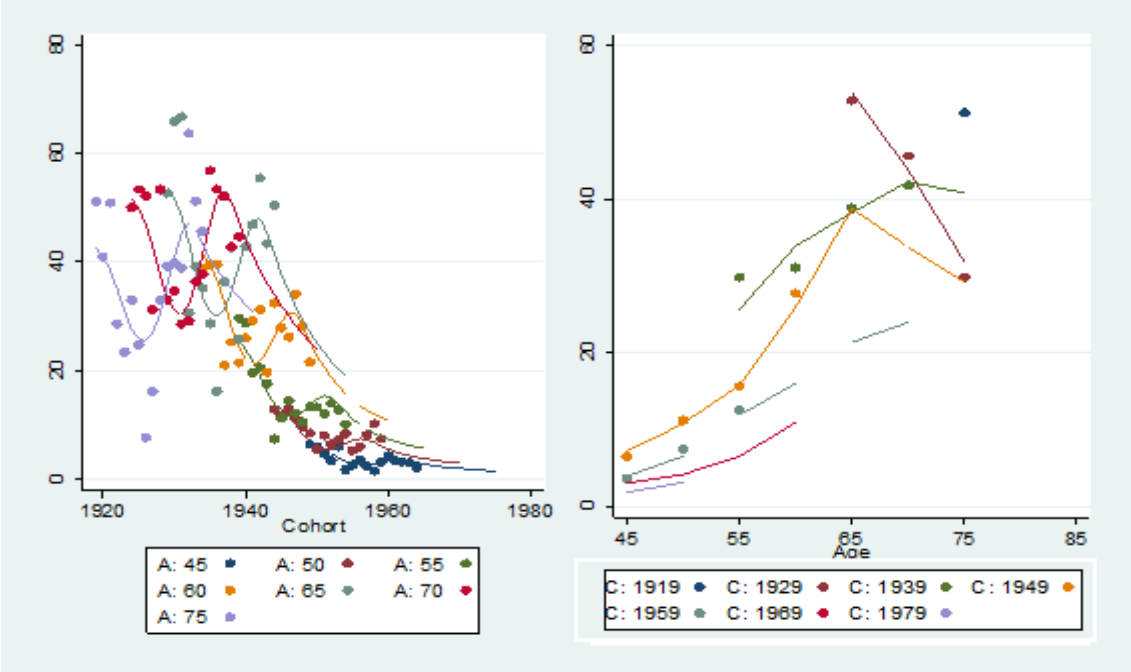


Figure 6.5: Actual (solid circles ••••) and fitted (solid curve) male cohort and age plots. In the left-hand panel, age-specific standardised rates are plotted against year of birth. In the right-hand panel, rates plotted against age and fitted values corresponding to different 10-year birth cohorts are joined together.

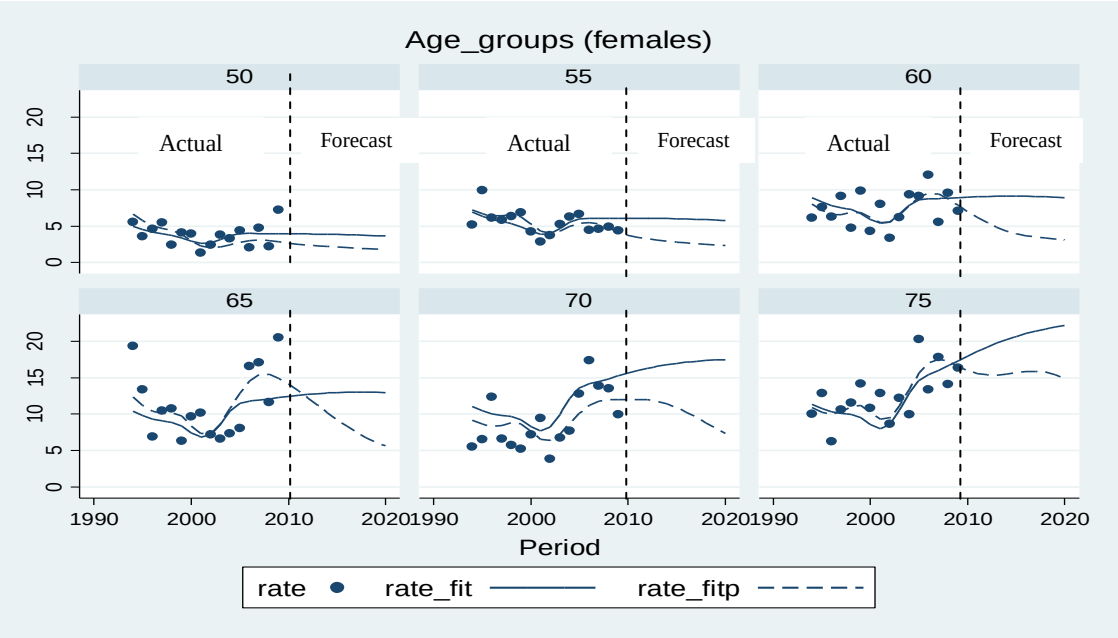


Figure 6.6: Actual (solid circles ••••) age-specific standardised rates of lung cancer incidence (per 100,000 person-year) with the fitted rate from 1994 to 2009 and the projected rate from 2010 to 2020 for females in KSA for age groups 50-75 years. Both the predictions based upon the logarithmic link (solid curve) and the predictions based on the power 0.2 link (dashed curve) are shown.

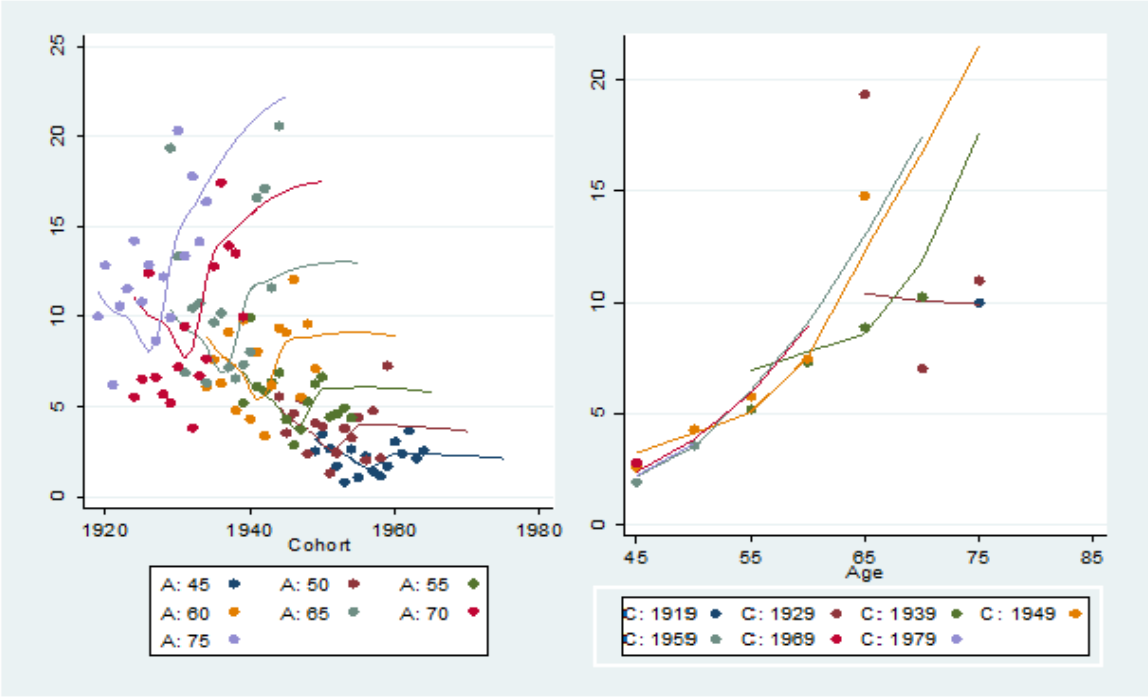


Figure 6.7: Actual (solid circles ••••) and fitted (solid curve) females cohort and age plots. In the left-hand panel, age-specific standardised rates are plotted against year of birth. In the right-hand panel, rates plotted against age and fitted values corresponding to different 10-year birth cohorts are joined together.

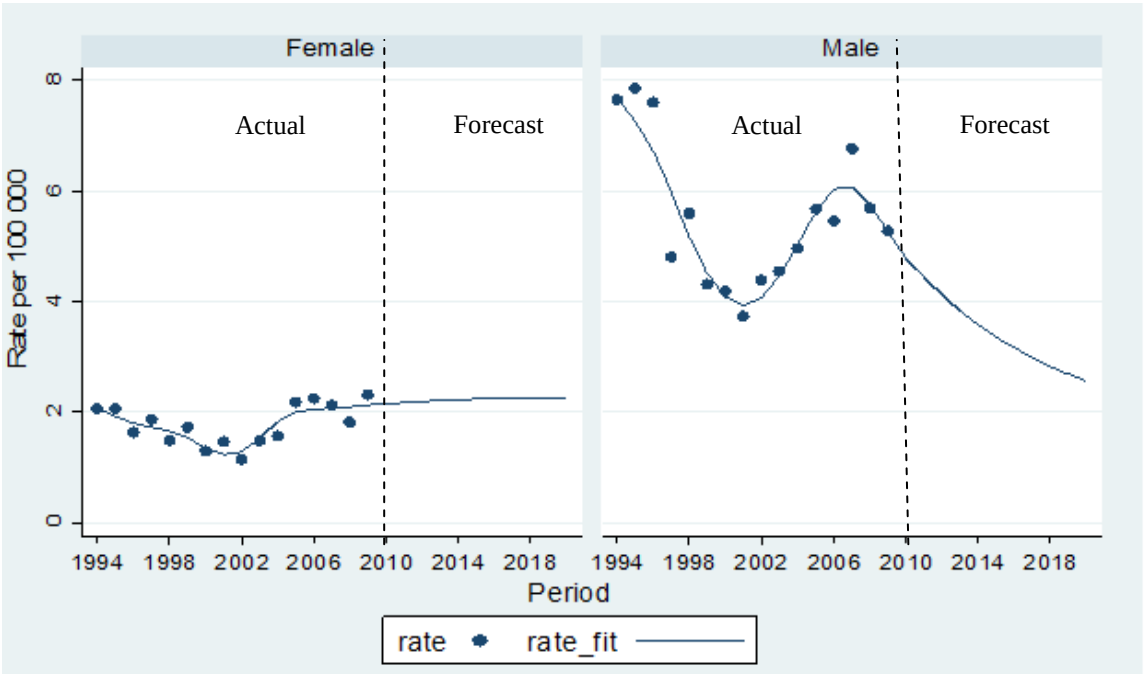


Figure 6.8: Actual (solid circles ••••) and fitted (solid curve) age standardised rates of lung cancer incidence in KSA (per 100,000 person-year) from 1994 to 2009 with forecast rates from 2010 to 2020 for males and females separately for age groups 0-75 years.

Trends in lung cancer incidence in Saudi Arabia are shown in Figure 6.8. Age-standardised incidence rates (ASR) for males lung cancer were at a minimum of 4 in 2001 with 209 cases per 100,000 whereas they were a maximum in 2007 at 6 with 380 cases per 100,000. Over the same time period females lung cancer incidence rates was minimum at 1.5 in 2001 with 56 cases per 100,000 whereas it was maximum in 2009 at 2.2 and 123 cases of lung cancer per 100,000 female population. The female ASR rate decreased gradually for 8 years and showed an upward increase until 2006. Thereafter, it maintained a steady increase. However, lung cancer incidence rates are still much lower in females than in males.

From Figure 6.9, lung cancer rate is projected to drop by approximately half between 1994 and 2020.

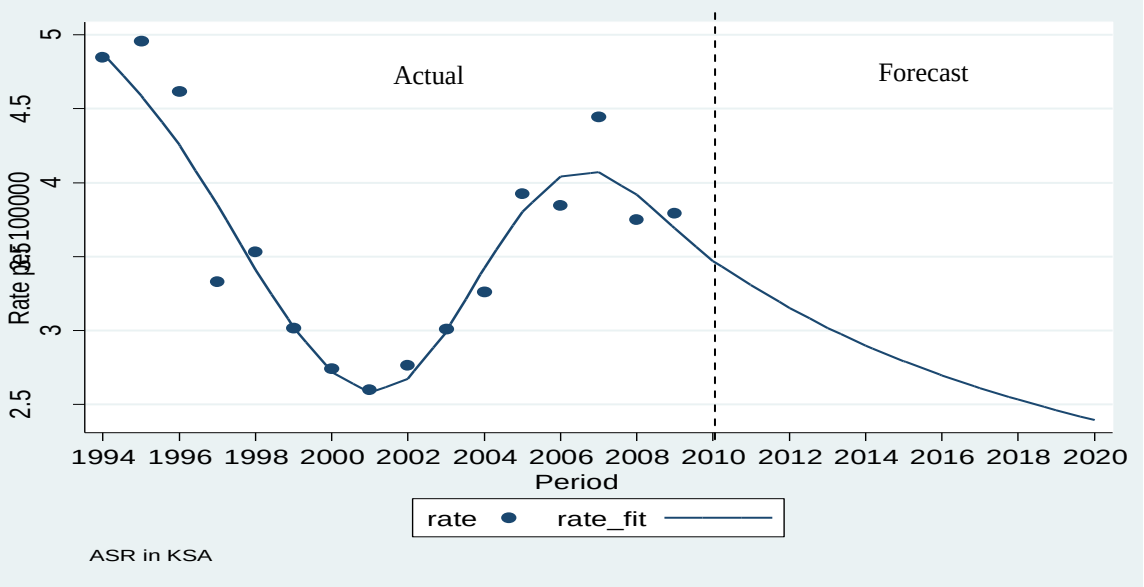


Figure 6.9: Actual (solid circles ••••) and fitted (solid curve) age standardised rate of lung cancer incidence in KSA for age groups 0-75 years (per 100,000 person-year) from 1994 to 2009 with forecast rate from 2010 to 2020.

The current age-specific incidence rates using the world standard population for lung cancer in Saudi Arabia are shown in Figure 6.10. In this graph, there are more cases of the disease diagnosed in males than in females. Figure 6.10 shows that lung cancer is rarely diagnosed in younger people before the age of 40 in KSA, but incidence rises sharply thereafter peaking in people aged 65-69 years. Most of the cases occur in people over the age of 50.

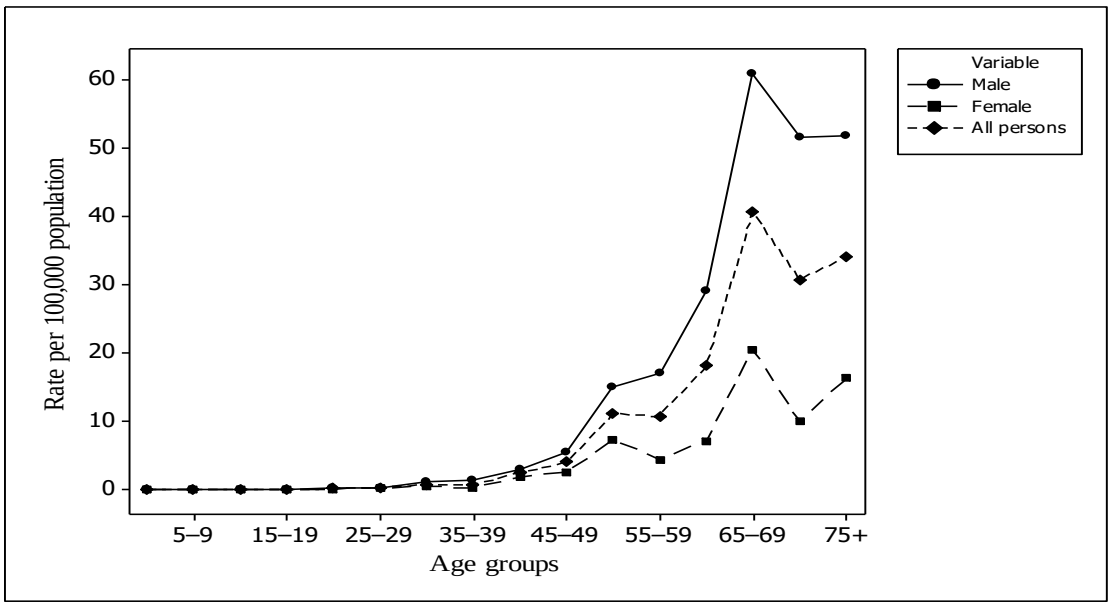


Figure 6.10: Age-specific incidence rates, lung cancer, by gender, KSA, 2009.

During the period from 1994 to 2009, an average of 300 cases per year and 85 cases per year were diagnosed in males and females respectively. This means that more cases were diagnosed in males than in females from 35-39 age-groups onwards in KSA (Figure 6.11). The projection in male cases indicates that there was a rise between 2009 and 2010, and then a sharp decline until 2015. Thereafter, the cases levelled up until 2020. In spite of this, female cases continued to increase gradually up to the year 2020.

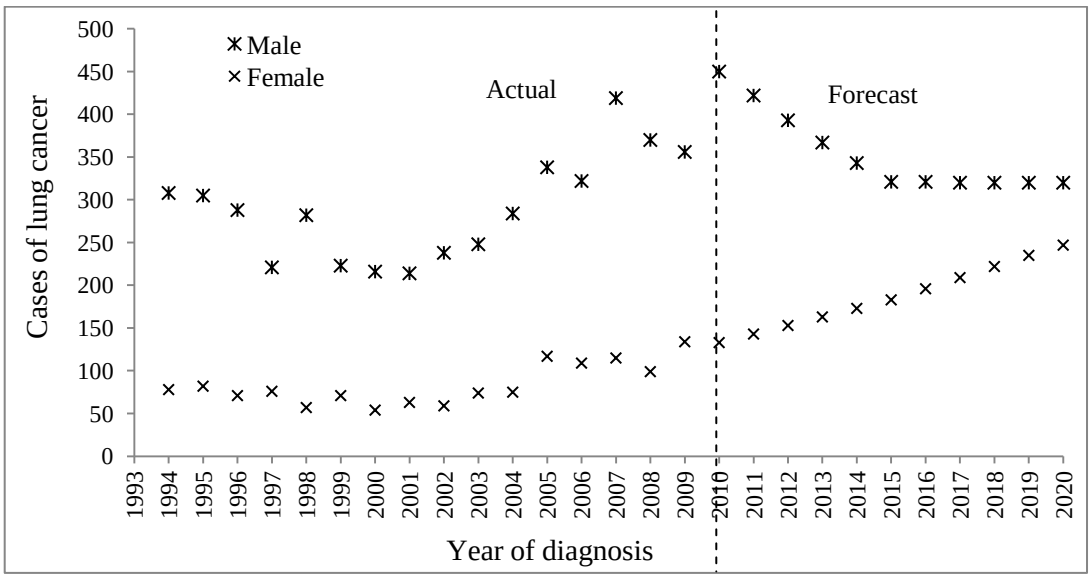


Figure 6.11: Number of new cases per year by gender in Saudi Arabia from 1994 to 2020.

6.7. Discussions

This study revealed that age, period, cohort, gender, ethnicity, and region effects are important factors for explaining lung cancer incidence rates in Saudi Arabia. We analyse the APC models in this chapter by using the restricted cubic splines to overcome the identification problem due to the exact linear relationship between age, period and cohort by fitting constraints to the model. In addition, the use of restricted cubic splines are useful because they produce a smooth overall curve and offer sufficient flexibility to capture the shape of most data, if appropriate knots are chosen.

The risk of lung cancer in males, adjusted by age, period, cohort, ethnicity and five regions, was approximately 79% greater than in females. Comparing the incidence rates of lung cancer among genders, it was found that although males show higher rates than females, females rate of lung cancer is expected to increase in the future. This is perhaps due to the increase in the proportion of female smokers.

The model selection statistics for example AIC, deviance and log likelihood were used to choose the best model of reduced and full three-way APC models with covariates. We conclude that none of the three components of the APC models should be eliminated from the model specification and selection. In addition, the covariates (with p-value shown) of gender (0.000), race (0.001), Southern region (0.044), Western region (0.030) and Eastern region (0.038) are statistically significant. Thus, the full APC model with the covariates of gender, race and five regions fit the data significantly better than the other four models (see Table 6.8).

Using APC models with restricted cubic splines, we are able to identify the effects of the three variables on the outcome of lung cancer incidence, which is the first step for exploring the causal processes of lung cancer. Having known that age is the most important time-related variable that influences the risk of lung cancer, our results show that age has a strong association with lung cancer incidence rates, suggesting age-related causes such as cumulative exposures of smoking over time may be the main reason for increasing lung cancer incidence in Saudi Arabia. The age effect shows that the incidence rate increases as the age increases, with wider credible interval width at younger age groups due to the heterogeneity in the data where there are sparse and zero counts associated with the fitted model (see Table F13 in Appendix F). For the period effect, the rate-ratio declined for about a decade to the early 2000s and then rose up to 2007, and thereafter observed a subsequent decrease, as an effect of the new polices implemented by the government during the period, such as the ban of smoking cigarettes in public places and the increase of tax on imported tobacco. The cohort effect reflects the cumulative effects of exposure in generations. Therefore, the cohort effect increased steadily up to the generation born before the second world war in 1939 and declined thereafter until the Gulf war in 1990. A subsequent increase followed in cohorts born after the Gulf war until 2009. However, our results show that the precision of the cohort effect was the lowest regarding to the widest credible intervals near the end of the cohort graph (see Figure 6.1). Particularly, the youngest cohort trends are uncertain due the low number of incidence cases. However, the complexity pattern of the cohort effect may be due to the short time period of the observed incidence data. Thus, more information is needed to further clarify our results.

Although the standard link function used in modelling the age-period-cohort models with covariates is the log, we used the power 0.2 link function for forecast purposes in line with the recommendations of Rutherford et al. (2012) and Sasieni (2012), because the log

link function tend to over-estimate future incidence rates. Therefore, we use the power 0.2 link function to dampen the exponential growth especially for long-term forecasts. The difference between the log and the power link functions can be seen clearly in female forecasts in Figure 6.6.

Our APC model provides a good fit to the incidence of lung cancer data compared to the A, P, AP, AC and PC models. This can be confirmed by our previous analysis in Chapter 4 of the best ARPDL(12,3,26,8) model when the model has shown that lung cancer is expected to decrease in males and increase in females. The APC model has provided good estimations for lung cancer forecast by using the fact that restricted cubic splines are linear beyond the boundary knot. The linear prediction beyond the range of the data was dictated by the shape of the data towards the end of the observation period ensuring that the forecasts give increased weight to more recent trends than standard approaches (Rutherford et al. 2012). Thus, based on the assumption that past period, cohort and age trends would continue into the future, we forecast the next 10 years of lung cancer in Saudi Arabia between 2010 and 2020. Our results show that ASR of lung cancer incidence is expected to decrease in males from 4.6 to 2.4 and increase in females from 2.0 to 2.2 per 100,000 population. This may be due to the increase of female smokers over time. The rate of lung cancer incidence in both genders is rarely diagnosed in younger age groups, but the rate rises sharply from people aged 65-69 years.

We forecast the rates and the cases of lung cancer incidence to 2020 to provide evidence for future policy making. However, we expect that the precision of our forecasts could be improved as further data are collected.

So far we have applied a range of alternative forecasting techniques to forecast lung cancer incidence in KSA from 1994 to 2009 with different data (monthly and yearly) with different covariates, to assess which method copes better with the specificities of each case. We use time series SARIMA modelling on monthly data for short-term forecasting. This appears to take into account trends and seasonal variations and eventually provide good estimates of current cases based on previous data. Also, we use generalised linear models such as APC modelling on yearly data for long-term forecasts and these appear to take into account the effects of age, period and cohort to extrapolate the future rate, which is important from the point of view of public health planers.

We describe a comparison of the quality of the forecasts generated by the APC model with classical time series SARIMA model. APC models can take one or more of the following into account, for example, population growth, ageing of the population and

changes in the rates based on the past observation. In addition, the advantages of using APC models take into account the age, period and cohort effects when forecasting future rates. On the other hand, the use of time series models are useful only for short-term forecast and can also explain the casual relationship between dependent and independent variables.

However, it is difficult to choose between SARIMA and APC models because these models may produce equally good fits to the data but offer different predictions. Predictions from APC models are uniquely determined (Holford 1985). Thus, we prefer APC models to the classical time series models because they extrapolate the effects of age, period and cohort into the future to make new forecasts. Most causes of lung cancer require prolonged exposure, determined by an aspect of life-style, such as smoking habits, which is fixed very early in adult life. In this case, a change in population exposure is more likely to manifest many years subsequently and will not occur simultaneously in all age groups; certain generations or cohort will have greater exposure than others and APC model will provide a better description of the data.

6.8. Summary

The observed data used for age-period-cohort (APC) modelling were annual incidence cases of lung cancer, for Saudi Arabia, by gender, ethnicity (race) and 5-year age group from 1994-2009. Incidence figures as mentioned in section 3.5, the total number of cases excluded from the ethnicity or race when dealing with the covariates includes 241 cases because of unknown nationalities. However, the overall model used the total number of lung cancer incident cases for both males and females. So this exclusion of ethnicity do not affect our forecasts.

It is often recommended to take the square root when one has a count data. On the other hand, when fitting a generalized linear model with a response variable distributed as Poisson (as in the APC approach), the log link is the canonical link. The log link implies a log transformation of the mean, λ, the parameter that governs the response distribution but not especially of a Poisson data. We gather the square root is best for stabilizing the variance and normalizing the Poisson distribution, and could have been considered for the earlier ARIMA/SARIMA models and in the distributed lag models.

The projection of the future rate depends on population projections; we use the 2010 to 2020 United Nations forecasts. Additionally, projection of lung cancer does not take into

account potential changes in lifestyle or treatment that could alter future rates of lung cancer incidence.

The forecast of lung cancer incidence presented in this chapter is based on the classical APC model, with the use of restricted cubic splines. In addition, the power link function is used instead of the logarithm link function to improve the forecast accuracy because the logarithm link function tend to over-estimate future incidence rates. Therefore, we use the power link function to reduce the exponential growth for long-term forecasts.

The estimated rates from APC modelling show a gradual decrease in males and a slight increase in females over the next 10 years. This is perhaps due to the increase in the proportion of female smokers. Male age standardised rates (ASR) of lung cancer are projected to fall to 2.4 per 100,000 by 2020, whereas female age standardised rates (ASR) of lung cancer are projected to increase to 2.2 per 100,000 by 2020. The growing and ageing populations will have a substantial impact, therefore the cases are projected to decrease in males (from 356 to 320) and to increase in females (from 134 to 247) between 2009 and 2020.

The results show that in Saudi Arabia, males have about a 79% greater incidence of lung cancer than females across the entire dataset when adjusting for the other effects. The p-value for the gender term highlights that the effect for gender is significant at the 0.1% level. In addition, the p-values for the covariates of race, Southern, Western, and Eastern regions show that the effects for these covariates are statistically significant.

Notwithstanding new potential changes in lifestyle or treatment, the incident cases of lung cancer in Saudi Arabia will decrease gradually in males reflecting the decrease of smoking prevalence among males and will increase slightly in females reflecting the ageing, growing populations and the increase of smoking prevalence in females.

CHAPTER 7

PREDICTION OF LUNG CANCER MORTALITY IN SAUDI ARABIA USING BAYESIAN DYNAMIC APC MODELLING

7.1. Introduction

Statistics provides analyses based on processing real data. Such analyses are noted for high measure of objectiveness, and thus provide information for making well informed decisions. Sometimes it is a great problem to gather enough data to describe the whole population.

Bayesian statistics is an effective tool for solving some inference problems when the available sample is too small for more complex statistical analysis to be applied. The lack of information may be offset (up to a certain point) by using Bayesian approach, as it enables us to utilise more sources of information.

In the Bayesian paradigm, we follow the strategy proposed by Held and Rainer (2001) and Shuichi et al. (2008) by using a dynamic age-period-cohort model to smooth age, period and cohort trends and to extrapolate N future periods and cohorts. Broadly, the methodology of the model building is a Bayesian version of the APC as suggested by Berzuini et al. (1993) and Besag et al. (1995). Bayesian dynamic APC modelling is expected to smooth the effect of age, period and cohort as much as possible in order to minimize the error and improve the predictions. By comparing the classical APC formulations to Bayesian APC, the predictions based on Bayesian APC do not rely on strong parametric assumptions for future values of subjective cohort and period effects and therefore seem to be particularly well suited for our objective. In addition, the models can take any additional unstructured heterogeneity. For more information on Bayesian APC models see the following papers (Berzuini et al, 1993; Berzuini and Clayton, 1994; Besag et al, 1995; Bray et al, 2001; Knorr-Held and Rainer, 2001; Bray, 2002; Baker and Bray, 2005; Schmid and Held, 2007).

This chapter is organized as follows. We give an overview of APC models and autoregressive models and introduce our dynamic Bayesian APC model in Section 7.2. We give some details on implementation and projection issues to these models in Section 7.3. For more information on the implemented models, see Appendix C. Section 7.4 outlines the analysis of the KSA lung cancer mortality data in three separate steps. The first step is an analysis of the complete data without any projection. We then conducted sensitivity analysis using four different values for the prior standard deviations of the age, period and

cohort effects, to evaluate the robustness of the results. Finally, we present a practical example of combined male and female lung cancer mortality modelling with forecasts until the year 2020. To conclude, we summarize our findings and propose next steps for research in section 7.6.

7.2. The Bayesian APC Model

Although the classical APC modelling produces almost the same results in estimating cancer rates as in Bayesian APC models, Bayesian APC models provide more robust results especially when the data are sparse (a lot of zero counts) (Raifu and Arbyn, 2009). However, Bayesian approaches are more complex and time consuming for researchers. Bayesian APC has been used more frequently in the last few years in epidemiology, demography, social & political behaviour and cancer research to predict cancer incidence and mortality rates (Baker and Bray 2005; Raifu and Arbyn 2009). Moreover, Bayesian APC models are recommended recently because it allows the uncertainty associated with functions of the parameters to be readily explored (Cleries et al., 2010).

The Bayesian approach considers the likelihood for the data and a prior belief about the smoothness of the model parameters. To obtain the posterior distribution, the model is constructed and simulated through the Markov Chain Monte Carlo (MCMC) method using Gibbs sampling. Then the best model is selected based on one of the goodness-of-fit criteria (Kaplan, 2014). Thus, the posterior distribution of μ is summarised as

$$p(\mu|y) = p(y|\mu)\, p(\mu)/p(y)$$

where = $p(y|\mu)$ is the likelihood function, $p(\mu)$ is the prior distribution of μ before seeing the data and $p(y)$ is the marginal distribution of the data.

APC models were originally proposed by sociologists and demographers in the early 1970s, see for example, (Mason et al., 1973). Bayesian APC approaches have been proposed firstly by Berzuini et al., (1993), Berzuini and Clayton (1994), and Besag et al., (1995). To smooth the prior of the model parameters, several methods have been proposed during the last 30 years, in such a way that the identification issue is avoided, for more details see Chapter 2 in page 24. This mean that improper priors could generate problems in making inference. Therefore, prior distributions should be selected carefully based on previous studies in the literature or on subjective prior beliefs.

7.3. Dynamic Age-period-cohort Model

Let $i(i = 1, \dots, I)$ index the age groups, where age group 1 includes 25-29 year olds, age group 2 includes 30-34 year olds, and so on; $j(1, \dots, J)$ index 1 year period, with period 1 as 1994, period 2 as 1995, and so on; and $k(1, \dots, K)$ index cohort. In our dataset, $I = 11$, $J = 16$ and $K = 66$. The following assumptions were made during the construction of the model.

The number of deaths in age group i, period j and cohort k is denoted Y_{ijk}, and is a realisation of Poisson random variable with mean λ_{ijk}, where

$$\log(\lambda_{ijk}) = \log(n_{ijk}) + \alpha_i + \beta_j + \gamma_k.$$

Here α_i, β_j, and γ_k are the effects of age group i, time period j and birth cohort k. The size of the population at risk, assumed to be known without error from census data, is denoted as n_{ijk}, and was used to transform the raw cases in both Table F15 and F16 (see Appendix F) to the rates in Table 3.7. As mentioned earlier in the incidence case, inclusion of the offsets $\mathrm{n_{ijk}}$ in the model for the Poisson mean implies that we are effectively modelling mortality rates $\lambda_{\mathrm{ijk}}/\mathrm{n_{ijk}}$, thereby correcting for the number at risk. It is clear that the parameterization is not identifiable, as we are using three co-ordinates to index into a two dimensional table of counts. In particular, $\mathrm{k} = 11 - \mathrm{i} + \mathrm{j}$. This methodological challenge results from the exact linear relationship between age, period, and (birth) cohort: cohort = period - age. Consequently, it is impossible to obtain valid estimations of the distinct effects of age, period, and cohort from standard regression-type models.

7.3.1. Prior Distributions for Age, Period and Cohort Effects

The prior distribution used in this analysis is a non-informative uniform distribution because we want the hyper-parameters to be estimated mainly from the data. If we are able to estimate the prior correctly, then the posterior mean will lie between the prior and the likelihood parameters. On the other hand, if we do not have information about the prior then the posterior parameters will be approximately the same as the maximum likelihood parameters and the effects of age, period and cohort will be close to maximum likelihood estimates (Congdon, 2006; Kaplan ,2014, pp 33-40).

In this Bayesian analysis, trends were modelled by using specific smoothing of model parameters because the cases of lung cancer mortality data are low. A 2nd order random walk (RW2) constraint has been used for age, period and cohort effects (Knorr-Held and Rainer, 2001), whilst 2nd order differences of this RW2 have been constrained for age

parameters, assuming that one 2nd order difference is estimated as the mean value on the previous and subsequent 2nd order differences (Cleries et al., 2006). Supposeα, β and γ are the age, period, and cohort effects respectively. Therefore, the age effect is constrained to

$$\alpha_i|\alpha_j\ , j\neq i \sim N(\mu_{\alpha_i|\alpha_j,j\neq i}, \tau_a)$$

$$\mu_{\alpha_1|\alpha_2,\alpha_3} = 2\alpha_2 - \alpha_3$$

$$\mu_{\alpha_2|\alpha_1,\alpha_3,\alpha_4} = \frac{2\alpha_1+4\alpha_3-\alpha_4}{5}$$

$$\mu_{\alpha_i|\alpha_{i+1},\alpha_{i+2},\alpha_{i-1},\alpha_{i-2}} = \frac{4\alpha_{i-1}+4\alpha_{i+1}-\alpha_{i-2}-\alpha_{i+2}}{6}, 3 \leq i \leq A-2$$

$$\mu_{\alpha_{A-1}|\alpha_A,\alpha_{A-2},\alpha_{A-3}} = \frac{2\alpha_A+4\alpha_{A-2}-\alpha_{A-3}}{5}$$

$$\mu_{\alpha_A|\alpha_{A-1},\alpha_{A-2}} = 2\alpha_{A-1} - \alpha_{A-2}$$

$$\tau_a = K_s\frac{1}{\sigma_a^2}$$

σ_a ~Uniform (0.01, 1)

where α_i is the effect of the ith age group (1,...,A), $\mu_{\alpha_i|\alpha_j,j\neq i}$ is the mean age effect for an individual aged i in the smooth prior specification, σ_a is the prior standard deviation and τ_a is the prior precision (inverse of the prior variance). It is advised that for hierarchical models, the prior standard deviation of the parameters should be modelled using non-informative uniform distributions on the interval [0.01, 1] which they are expected to improve the estimations especially when the variables are below five (Gelman, 2005). An adaptive precision parameter is denoted as K_s and has been assumed the same for age, period, and cohort effects. The period parameters, $\{\beta_1, \dots, \beta_P\}$, have been modelled using RW2 as follows:

$$\beta_1 = 0$$

$$\beta_2 \sim N(0, \tau_p)$$

$$\beta_i|\beta_{i-1},\beta_{i-2} \sim N\left(\mu_{\beta_i|\beta_{i-1},\beta_{i-2}},\ \tau_p\right), 3 \leq i \leq P$$

$$\mu_{\beta_i|\beta_{i-1},\beta_{i-2}} = 2\beta_{i-1} - \beta_{i-2}, 3 \leq i \leq P$$

$$\tau_p = K_s\frac{1}{\sigma_p^2}$$

σ_p ~Uniform (0.01, 1)

Notice that $\beta_1 = 0$ and so the first period is the reference period.

Similarly, cohort parameters, $\{\gamma_1, \dots, \gamma_C\}$, were modelled through

$$\gamma_1 = 0$$

$$\gamma_2 \sim N(0, \tau_c)$$

$$\gamma_i|\gamma_{i-1},\gamma_{i-2} \sim N\left(\mu_{\gamma_i|\gamma_{i-1},\gamma_{i-2}},\tau_c\right), 3 \le i \le C$$

$$\mu_{\gamma_i|\gamma_{i-1},\gamma_{i-2}} = 2\gamma_{i-1} - \gamma_{i-2}, 3 \le i \le C$$

$$\tau_c = K_s \frac{1}{\sigma_c^2}$$

σ_c ~Uniform (0.01, 1)

Notice that $\gamma_1 = 0$ and so the first cohort is the reference cohort.

7.4. Materials and Methods

The mortality rates of lung cancer were calculated using the population of Saudi Arabia according to the statistical national census of 1994 to 2009. The rates were age standardised using the world standard population. It was decided to restrict the age range between 25-75 because the observation number of lung cancer mortality is low in the earliest age groups and this might lead to less precision in the estimates. Thus, data were arranged in one-year interval period from 1994 to 2009 and 5-year age group from 25-29 years to 75+ years. The periods and the age groups involved 66 (5. (I - 1) + J) overlapping 5-year cohorts (Held and Rainer, 2001). The data are provided in Appendix D. The cohort groups started from 1919 cohort and finishing with the cohort 1980. The form of the model falls into the class of generalized linear models to assess the effects of the three variables assuming that the number of lung cancer mortality follows a Poisson distribution. Three models were estimated, namely, APC, AP and AC models. Comparison between nested models was evaluated by the changes in Deviance Information Criterion DIC. The best-fitting model is chosen by the lowest value of DIC. Bayesian dynamic APC model smoothing and Markov Chain Monte Carlo (MCMC) techniques were used. Constraints on 2nd order differences were used for all the three effects. Additionally, the posterior inference were based on 2500, 5000, 10000, 50000 and 100000 iterations of Gibbs sampler after a burn-in of 1000 iterations was discarded. Convergence was assessed by using the Gelman and Rubin diagnostic statistic. R and R2WinBUGS statistical software were used for the implementation. In addition, a second order random walk (RW2) has been used for age, period and cohort effects to smooth the models as possible. Furthermore, we introduced an adaptive precision parameter (K_s) for each prior distribution of age, period and cohort to smooth the parameter effects as much as possible as suggested by Cleries et al. (2010). The models have been fitted for age-period (AP) , age-cohort (AC) and age-period-cohort (APC) and the one with lowest value of deviance information criteria (DIC)

is selected as the best model. DIC can be calculated directly by adding the number of model parameters (pD) to the posterior deviance.

Spiegelhalter et al. (2002) has proposed a method for judging the goodness-of-fit for Bayesian model comparison. The criterion is based on the deviance given by the following formula

$$D(\mu) = -2\,log\{\,p(y|\mu)\} + 2log\,\{f(y)\},$$

where $D(\mu)$ is the posterior mean and $f(y)$ is some fully specified standardizing term that is a function of the data alone. Thus, the DIC is given by

$$DIC = \overline{D(\mu)} + pD$$

where $\overline{D(\mu)}$ is the posterior expectation of the deviance and pD is the effective number of model parameters.

7.5. Results

7.5.1. Bayesian Model Comparison and Sensitivity Analysis

Comparison between nested models was evaluated by the changes in Deviance Information Criterion (DIC). The best-fitting model is chosen by the lowest value of DIC (Spiegelhalter, 2002). We evaluated the DIC values for AP, AC and APC models in five scenarios, depending on the value for the adaptive precision parameter, K_s. In this Bayesian analysis, we have evaluated K_s at five different fixed values (1, 0.1, 0.01, 0.001, 0.0001). The adaptive precision parameter selected has been used for the APC model used for projections.

We also carried out sensitivity analysis on the prior standard deviations of the age, period and cohort effects, to evaluate the robustness of the results. In particular, we run the simulations for the prior standard deviations for the three effects at four different fixed values (1.0, 0.25, 0.5 and 0.75) each with AP, AC, and APC models. The number of iterations were simultaneously altered between 2500 and 100000 after a burn-in of 1000 iterations was discarded for each single model to check the convergence. Results of the various simulations considered in this analysis have been tabulated as follows.

7.5.2. Sensitivity Analysis for the Best Bayesian AP Model

Tables from 7.1 to 7.4 show different values of DIC and pD with different values of the adaptive precision parameter $K_s \in \{1, 0.1, 0.01, 0.001, 0.0001\}$ for the reduced AP model at different fixed prior standard deviations of 1, 0.25, 0.5 and 0.75. These tables present the procedure for the selection of the adaptive precision parameter K_s using Deviance Information Criterion (DIC) and the effective number of model parameters (pD). Through the tables, the influence of the adaptive precision parameters on the predictive performance of the models due to model with lowest DIC value best predictive performance shows. In this analysis, we have selected the value 0.001 because AP model tested showed the lowest DIC value and this was used for the APC model used for predictions. To conclude, the adaptive precision parameter selected due to DIC value is the minimum observed among models. Hence, small adaptive precision parameter implies small variance of the age and period effects.

Table 7.2 shows the best age-period model when the adaptive precision parameter is 0.001 and the prior standard deviation is 0.25 regarding the lowest value of DIC and the stabilization of the iterations as it increases.

Table 7.1: Different values of DIC and pD with different values of the adaptive precision parameter for the AP model when the prior standard deviation is 1.0.

Iteration	K_s= 1.0		K_s= 0.1		K_s= 0.01		K_s= 0.001		K_s= 0.0001	
	DIC	pD	DIC	pD	DIC	pD	DIC	pD	DIC	pD
2500	305.04	11.13	301.26	10.79	299.67	11.97	299.17	14.60	305.18	20.06
5000	302.65	11.31	301.80	11.68	300.12	11.92	299.24	14.91	305.70	20.59
10000	303.92	11.37	302.66	11.60	300.75	12.01	299.50	15.03	305.51	20.54
50000	303.02	11.70	303.28	11.92	300.64	12.11	299.97	15.09	305.57	20.46
100000	302.80	11.76	302.88	11.76	300.19	12.16	299.57	15.01	305.58	20.42

Table 7.2: Different values of DIC and pD with different values of the adaptive precision parameter for the AP model when the prior standard deviation is 0.25.

Iteration	K_s= 1.0		K_s= 0.1		K_s= 0.01		K_s= 0.001*		K_s= 0.0001	
	DIC	pD	DIC	pD	DIC	pD	DIC	pD	DIC	pD
2500	302.24	11.64	302.14	11.52	418.81	13.21	299.31	15.09	303.18	19.94
5000	304.22	11.18	302.90	12.69	301.26	12.35	298.84	14.91	304.45	20.24
10000	303.45	11.57	302.69	11.71	300.68	12.02	299.01	14.90	304.88	20.22
50000	303.14	11.43	302.97	11.72	300.27	12.05	299.44	15.07	305.15	20.37
100000	303.77	11.55	302.72	11.74	300.54	12.18	299.31	15.02	305.31	20.39

*=Best adaptive precision parameter.

Table 7.3: Different values of DIC and pD with different values of the adaptive precision parameter for the AP model when the prior standard deviation is 0.50.

Iteration	K_s = 1.0		K_s = 0.1		K_s = 0.01		K_s = 0.001		K_s = 0.0001	
	DIC	pD	DIC	pD	DIC	pD	DIC	pD	DIC	pD
2500	302.03	11.49	303.19	11.62	303.87	12.06	299.31	14.70	315.41	20.81
5000	304.54	11.51	303.06	11.66	299.32	11.97	299.46	15.01	305.77	20.92
10000	322.89	13.20	302.52	11.80	299.32	12.00	299.76	15.07	305.58	20.60
50000	302.94	11.48	302.54	11.57	300.27	12.25	299.79	15.06	305.41	20.47
100000	303.38	11.48	302.35	11.71	301.59	12.45	299.67	15.08	305.43	20.45

Table 7.4: Different values of DIC and pD with different values of the adaptive precision parameter for the AP model when the prior standard deviation is 0.75.

Iteration	K_s = 1.0		K_s = 0.1		K_s = 0.01		K_s = 0.001		K_s = 0.0001	
	DIC	pD	DIC	pD	DIC	pD	DIC	pD	DIC	pD
2500	304.71	11.69	303.89	10.95	300.22	12.23	299.01	15.43	305.45	20.08
5000	302.86	11.35	302.08	11.85	300.73	12.00	299.78	15.01	304.44	20.42
10000	301.98	10.50	303.44	11.74	300.30	12.05	299.22	15.11	305.07	20.36
50000	302.76	11.86	302.64	11.57	300.16	12.06	299.43	15.07	305.35	20.39
100000	302.86	11.59	302.86	11.74	300.64	12.07	299.48	15.08	305.45	20.41

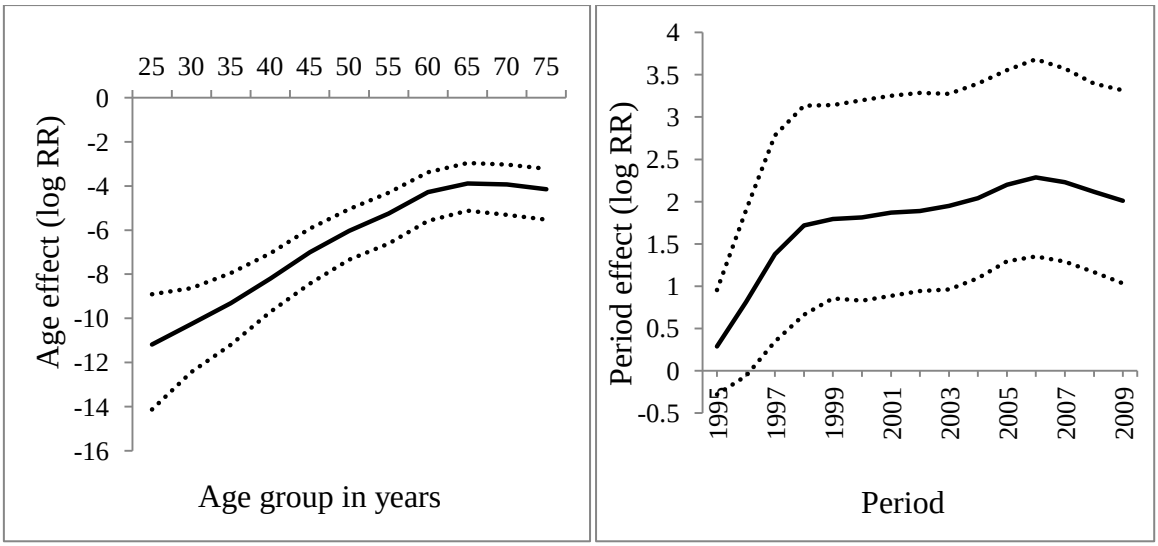


Figure 7.1: Effects of age and period on mortality from lung cancer identified by the age-period model for persons aged 25 to 75 years in Saudi Arabia during the period 1994-2009 within 95% credible intervals (dash lines).

7.5.3. Sensitivity Analysis for the Best Bayesian AC Model

Tables from 7.5 to 7.8 show different values of DIC and pD with different values of the adaptive precision parameter $K_s \in \{1, 0.1, 0.01, 0.001, 0.0001\}$ for the reduced AC model at different fixed prior standard deviations of 1, 0.25, 0.5 and 0.75. These tables show the procedure for the selection of the adaptive precision parameter K_s using Deviance Information Criterion (DIC) and the effective number of model parameters (pD). Notice that the influence of the adaptive precision parameters on the predictive performance of the models due to model with lowest DIC value best predictive performance shows. In this analysis, we have selected the value 0.1 because AC model tested showed the lowest DIC value and this was used for the APC model used for predictions. To conclude, the adaptive precision parameter selected due to DIC value is the minimum observed among models.

Table 7.8 shows the best age-cohort model when the adaptive precision parameter is 0.1 and the prior standard deviation is 0.75 regarding the lowest value of DIC. However, the DIC obtained here is far away from the DIC obtained from age-period model. Thus, it has been plotted in Figure 7.2 for just seek of comparison.

Table 7.5: Different values of DIC and pD with different values of the adaptive precision parameter for the AC model when the prior standard deviation is 1.0.

Iteration	K_s= 1.0		K_s= 0.1		K_s= 0.01		K_s= 0.001		K_s= 0.0001	
	DIC	pD	DIC	pD	DIC	pD	DIC	pD	DIC	pD
2500	335.43	10.34	323.78	10.90	321.63	15.50	322.91	21.69	343.95	36.20
5000	342.64	11.92	328.27	11.83	321.02	15.72	324.08	23.90	342.84	36.12
10000	328.06	11.40	317.60	11.29	319.11	16.61	324.03	23.93	342.12	35.79
50000	319.45	11.59	314.30	12.44	315.17	16.04	322.80	22.93	342.17	35.93
100000	312.91	11.64	312.43	12.49	314.33	16.11	322.84	23.12	366.91	36.23

Table 7.6: Different values of DIC and pD with different values of the adaptive precision parameter for the AC model when the prior standard deviation is 0.25.

Iteration	K_s= 1.0		K_s= 0.1		K_s= 0.01		K_s= 0.001		K_s= 0.0001	
	DIC	pD	DIC	pD	DIC	pD	DIC	pD	DIC	pD
2500	329.60	11.76	331.17	11.73	321.12	15.02	324.07	22.30	342.37	35.55
5000	333.57	9.71	324.85	12.01	323.72	15.60	323.66	22.30	342.65	35.76
10000	331.76	10.60	318.83	11.49	315.90	15.64	322.56	22.93	343.14	36.12
50000	324.03	10.79	315.00	12.86	314.47	15.76	323.48	23.35	342.27	35.88
100000	314.78	12.11	312.60	12.68	313.52	15.82	327.61	23.27	341.88	35.77

Table 7.7: Different values of DIC and pD with different values of the adaptive precision parameter for the AC model when the prior standard deviation is 0.50.

Iteration	K_s= 1.0		K_s= 0.1		K_s= 0.01		K_s= 0.001		K_s= 0.0001	
	DIC	pD	DIC	pD	DIC	pD	DIC	pD	DIC	pD
2500	331.63	11.45	323.53	12.19	319.19	14.38	324.28	22.15	342.99	35.73
5000	330.51	13.66	323.71	11.48	321.86	15.26	366.09	23.49	343.07	35.94
10000	325.45	11.34	324.81	12.51	317.85	15.78	323.05	23.20	342.40	35.92
50000	320.90	11.49	313.49	12.09	314.73	15.69	322.95	23.04	342.24	35.82
100000	313.37	10.56	341.60	14.51	313.56	15.91	322.98	23.19	342.43	35.88

Table 7.8: Different values of DIC and pD with different values of the adaptive precision parameter for the AC model when the prior standard deviation is 0.75.

Iteration	K_s= 1.0		K_s= 0.1*		K_s= 0.01		K_s= 0.001		K_s= 0.0001	
	DIC	pD	DIC	pD	DIC	pD	DIC	pD	DIC	pD
2500	332.04	10.77	318.72	11.85	326.06	14.82	325.02	22.37	405.84	35.26
5000	325.27	11.29	323.36	12.95	317.41	15.51	323.24	22.46	364.10	36.77
10000	324.75	10.40	322.33	12.90	344.12	16.19	322.34	22.57	342.48	35.86
50000	319.58	11.52	316.55	12.43	315.04	16.12	322.21	23.05	342.11	35.85
100000	314.76	11.36	310.36	12.09	313.00	15.90	322.68	23.09	342.47	36.00

*=Best adaptive precision parameter

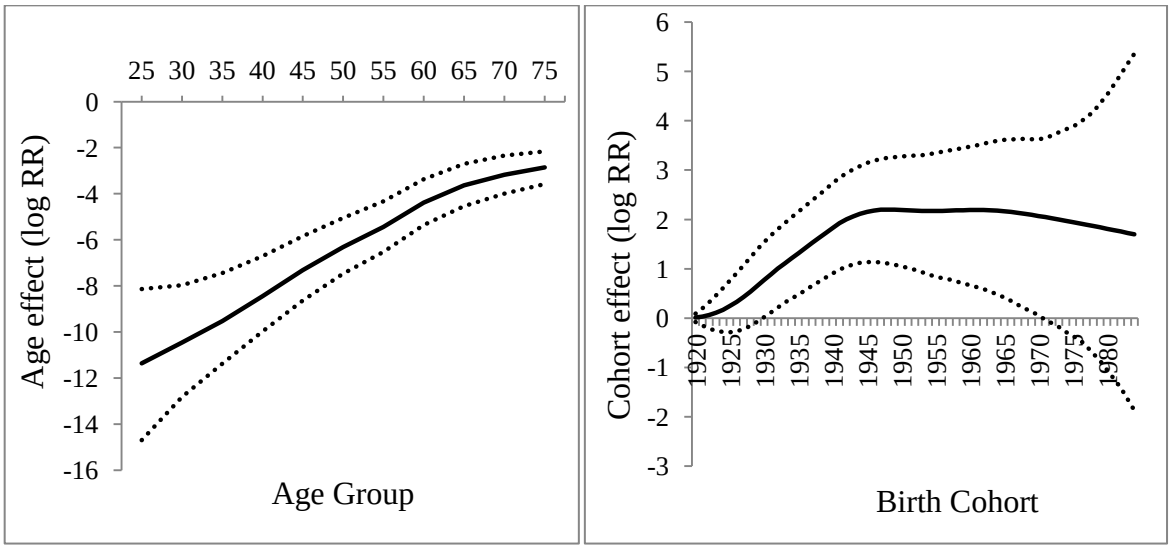


Figure 7.2: Effects of age and cohort on mortality from lung cancer identified by the age-cohort model for persons aged 25 to 75 years in Saudi Arabia during the period 1994-2009 within 95% credible intervals (dash lines).

7.5.4. Sensitivity Analysis for the Best Bayesian APC Model

Tables from 7.9 to 7.12 show different values of DIC and pD with different values of the adaptive precision parameter $K_s \in \{1, 0.1, 0.01, 0.001, 0.0001\}$ for the full APC model at different fixed prior standard deviations of 1, 0.25, 0.5 and 0.75. These tables present the procedure for the selection of the adaptive precision parameter K_s using Deviance Information Criterion (DIC) and the effective number of model parameters (pD). Note the influence of the adaptive precision parameters on the predictive performance of the models due to model with lowest DIC value best predictive performance shows. In this analysis, we have selected the value 1.0 because APC model tested showed the lowest DIC value and this was used for the APC model used for predictions. To conclude, the adaptive precision parameter selected due to DIC value is the minimum observed among models.

Table 7.10 shows the best age-period-cohort model when the adaptive precision parameter is 1.0 and the prior standard deviation is 0.25 regarding the lowest values of DIC. However, the DIC obtained here is not lower than the DIC obtained from age-period model. Thus, it has been plotted in Figure 7.3 for just seek of comparison.

Table 7.9: Different values of DIC and pD with different values of the adaptive precision parameter for the APC model when the prior standard deviation is 1.0.

Iteration	K_s= 1.0		K_s= 0.1		K_s= 0.01		K_s= 0.001		K_s= 0.0001	
	DIC	pD	DIC	pD	DIC	pD	DIC	pD	DIC	pD
2500	301.79	12.70	302.18	13.68	303.88	17.03	314.98	27.88	337.07	42.74
5000	301.82	12.98	301.76	14.27	306.39	18.95	316.20	28.76	337.34	42.83
10000	301.75	12.95	303.70	15.34	301.91	20.30	314.44	27.89	338.08	43.21
50000	301.15	13.91	301.18	14.88	305.50	19.53	314.79	28.23	338.21	43.48
100000	300.95	13.50	301.61	14.94	305.34	19.16	314.84	28.17	337.69	43.19

Table 7.10: Different values of DIC and pD with different values of the adaptive precision parameter for the APC model when the prior standard deviation is 0.25.

Iteration	K_s= 1.0*		K_s= 0.1		K_s= 0.01		K_s= 0.001		K_s= 0.0001	
	DIC	pD	DIC	pD	DIC	pD	DIC	pD	DIC	pD
2500	303.08	13.10	303.43	15.03	323.42	19.45	314.59	28.12	335.57	42.22
5000	299.94	13.10	300.67	14.12	304.10	18.66	314.06	27.89	338.07	43.26
10000	301.11	13.45	302.24	14.18	305.59	19.84	315.13	28.34	337.76	43.27
50000	300.52	12.72	302.57	15.16	305.41	19.43	315.19	28.19	337.96	43.28
100000	302.36	13.48	303.05	15.99	304.60	18.97	314.62	28.16	337.73	43.22

*=best adaptive precision parameter

Table 7.11: Different values of DIC and pD with different values of the adaptive precision parameter for the APC model when the prior standard deviation is 0.50.

Iteration	K_s= 1.0		K_s= 0.1		K_s= 0.01		K_s= 0.001		K_s= 0.0001	
	DIC	pD	DIC	pD	DIC	pD	DIC	pD	DIC	pD
2500	302.53	11.40	303.02	13.80	306.55	17.86	314.50	27.49	338.59	43.23
5000	303.81	13.74	303.32	15.07	306.43	18.55	315.25	28.17	337.78	43.29
10000	301.23	12.27	301.53	13.99	305.50	19.42	315.55	28.44	337.71	43.25
50000	301.18	13.08	302.52	15.36	304.94	19.00	314.61	28.17	337.99	43.32
100000	301.54	13.38	301.83	15.01	305.35	19.34	314.63	28.14	338.01	43.31

Table 7.12: Different values of DIC and pD with different values of the adaptive precision parameter for the APC model when the prior standard deviation is 0.75.

Iteration	K_s= 1.0		K_s= 0.1		K_s= 0.01		K_s= 0.001		K_s= 0.0001	
	DIC	pD	DIC	pD	DIC	pD	DIC	pD	DIC	pD
2500	305.08	12.30	304.15	13.19	307.02	19.04	315.72	28.36	338.83	44.07
5000	304.50	12.80	302.10	14.40	305.86	19.17	314.67	27.84	337.56	42.81
10000	301.75	14.08	302.43	14.28	305.46	19.28	315.04	27.94	339.40	43.55
50000	300.02	13.27	301.29	14.58	304.68	19.28	315.12	28.19	338.20	43.42
100000	301.66	13.34	302.51	15.37	305.00	19.17	314.87	28.22	337.91	43.34

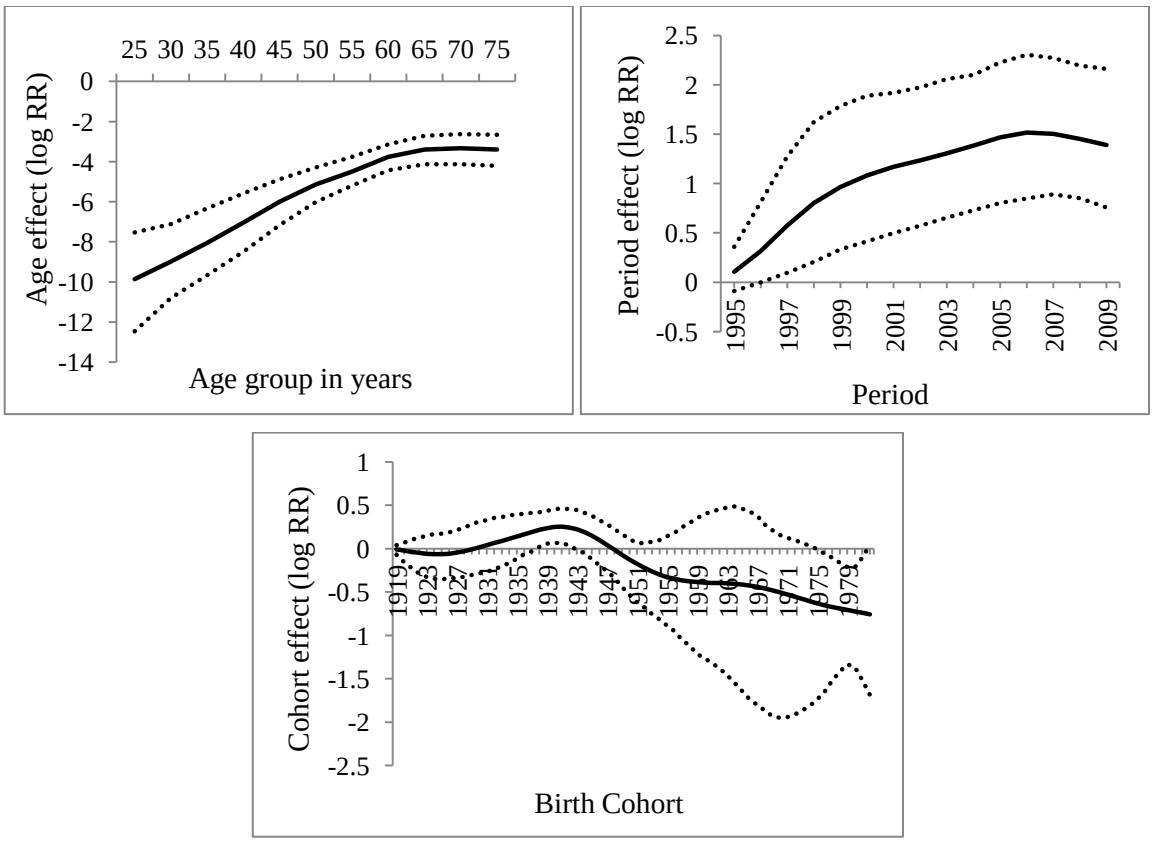


Figure 7.3: Effects of age, period and cohort on mortality from lung cancer identified by the age-period-cohort model for persons aged 25 to 75+ years in Saudi Arabia during the period 1994-2009 within 95% credible intervals (dash lines).

From above, for example, the full Bayesian APC model (Figure 7.3) illustrates the non-linear effects of age, period and cohort on lung cancer mortality for both gender combined. The age effect shows a dramatic increase of lung cancer mortality up to the age class 65-69 and then it starts to decrease gradually for the rest of age groups. The non-linear effects of the period of lung cancer mortality show again an increase of deaths up to the year 2007 and start to fall slightly. The cohort effects show a fluctuated and an increased pattern until it reached the peak on the birth cohort 1939 and since then it started to decrease until 2009.

Tables 7.1- 7.12 above show the criteria for selecting the adaptive precision parameter K_s. The influence of the adaptive precision parameters can be accessed through the different DIC values for AP, AC, and APC models with different values of prior standard deviations. From this Bayesian analysis, we have selected across the tables, the value of K_s =0.001 when the prior standard deviation is 0.25 because the age-period model tested showed the lowest DIC value (see Table 7.2). Notice that the parameters of the interval of the uniform distribution used in the best AP model is [0.01, 1]. Notwithstanding, we subjected again the best AP model to sensitivity analysis on the hyper-prior distribution to evaluate the robustness of the result. In particular, we changed the interval parameters of the uniform distribution for the age and period. The results of this analysis are provided in Table 7.13.

Table 7.13: Bayesian AP modelling using non-informative prior (uniform distribution) with varying intervals (endpoints).

	K_s=0.001 and prior standard deviation =0.25													
Iteration	(0.01,1)*		(0.1,100)		(0.001,1)		(0.1,0.5)		(0.01,0.5)		(0.01,0.75)		(0.01,0.25)	
	DIC	pD	DIC	pD	DIC	pD	DIC	pD	DIC	pD	DIC	pD	DIC	pD
2500	299.3	15.1	310.7	24.0	302.7	11.6	310.7	23.8	301.2	15.3	301.2	15.3	299.3	14.9
5000	298.8	14.9	310.4	24.0	302.5	12.0	310.5	23.7	300.1	14.8	300.8	15.1	300.1	15.0
10000	299.0	14.9	310.2	23.9	302.8	11.8	310.3	23.7	300.0	14.9	300.3	15.1	299.3	15.2
50000	299.4	15.1	310.2	23.8	303.0	11.7	310.4	23.6	299.3	15.0	299.4	15.0	299.2	15.1
100000	299.3	15.0	310.5	23.9	302.7	11.8	310.7	23.2	299.4	15.0	299.3	15.0	299.4	15.1

*=Best parameters.

From Table 7.13 above, the results obtained with new intervals did not show any improvement than the best result obtained so far.

Table 7.14: Summary Table of results. Overall best Bayesian APC model is stared.

Model	Prior standard deviation	K_s	DIC	pD
AP*	0.25	0.001	298.840	14.910
AC	0.75	0.100	310.360	12.090
APC	0.25	1.000	299.940	13.100

From Table 7.14, the adaptive precision parameter selected due to DIC value is the minimum observed among these models. Hence, small adaptive precision parameter implies small variance of the age and period effects.

7.5.5. Model Validation

One way to check if our chain has converged is to see how well our chain is mixing, or moving around the parameter space. The following figures show the graphical convergence diagnosis of the MCMC algorithms of selected parameters due to the limited space here. Thus, the first two parameters in Figure 7.4 represent the effects of the firs age group (25-29) and the first period (1994). For each selected parameter, the trace plot illustrates the posterior sample values of a parameter during the runtime of the chain and the marginal density plot is the smoothened histogram of the parameter values from the trace plot. Therefore, the trace plots provide evidence of satisfactory convergence of the MCMC algorithms for these two parameters. The last three parameters represent the deviance, prior standard deviation for age and period. The trace plots indicate each chain is mixing well for each single parameter. Additionally, the Gelman-Rubin convergence is used as a formal test for convergence that assesses whether parallel chains with dispersed initial values converge to the same target distribution. The Gelman-Rubin diagnostic demonstrates that the scale reduction factor for each parameter is equal to one indicating no difference between the chains for a particular parameter. The multivariate potential scale reduction factor is also one, suggesting the joint convergence of the chains over all the parameters. Gelman-Rubin diagnostic plots for selected parameters are presented in Figure 7.5. It can be seen that for each parameter, the Gelman-Rubin plots illustrate the development of Gelman-Rubin's shrink factor as the number of iterations increases and the shrink factor of each parameter eventually stabilized around one.

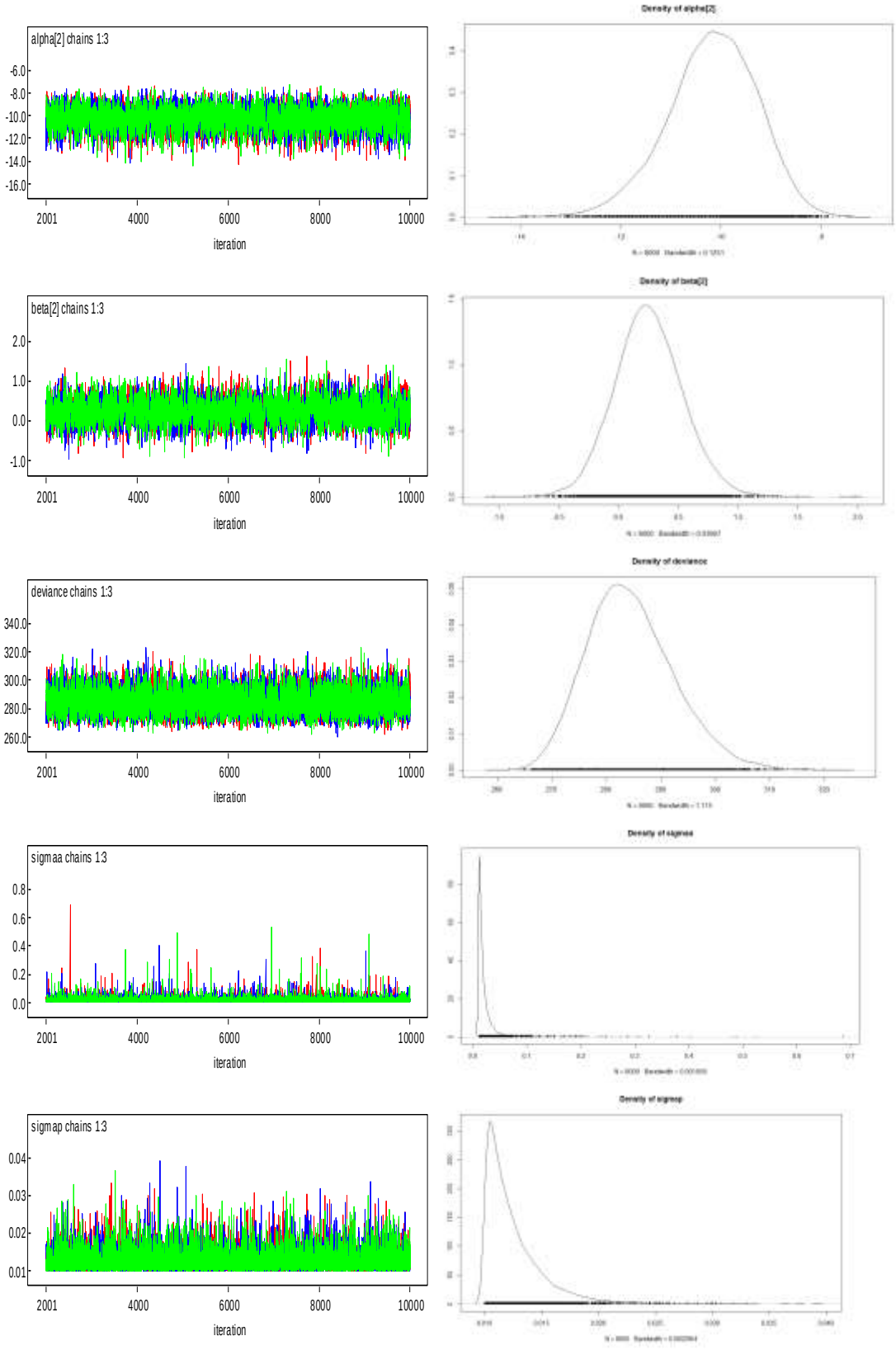


Figure 7.4. Trace and density plots for the posterior samples of selected parameters.

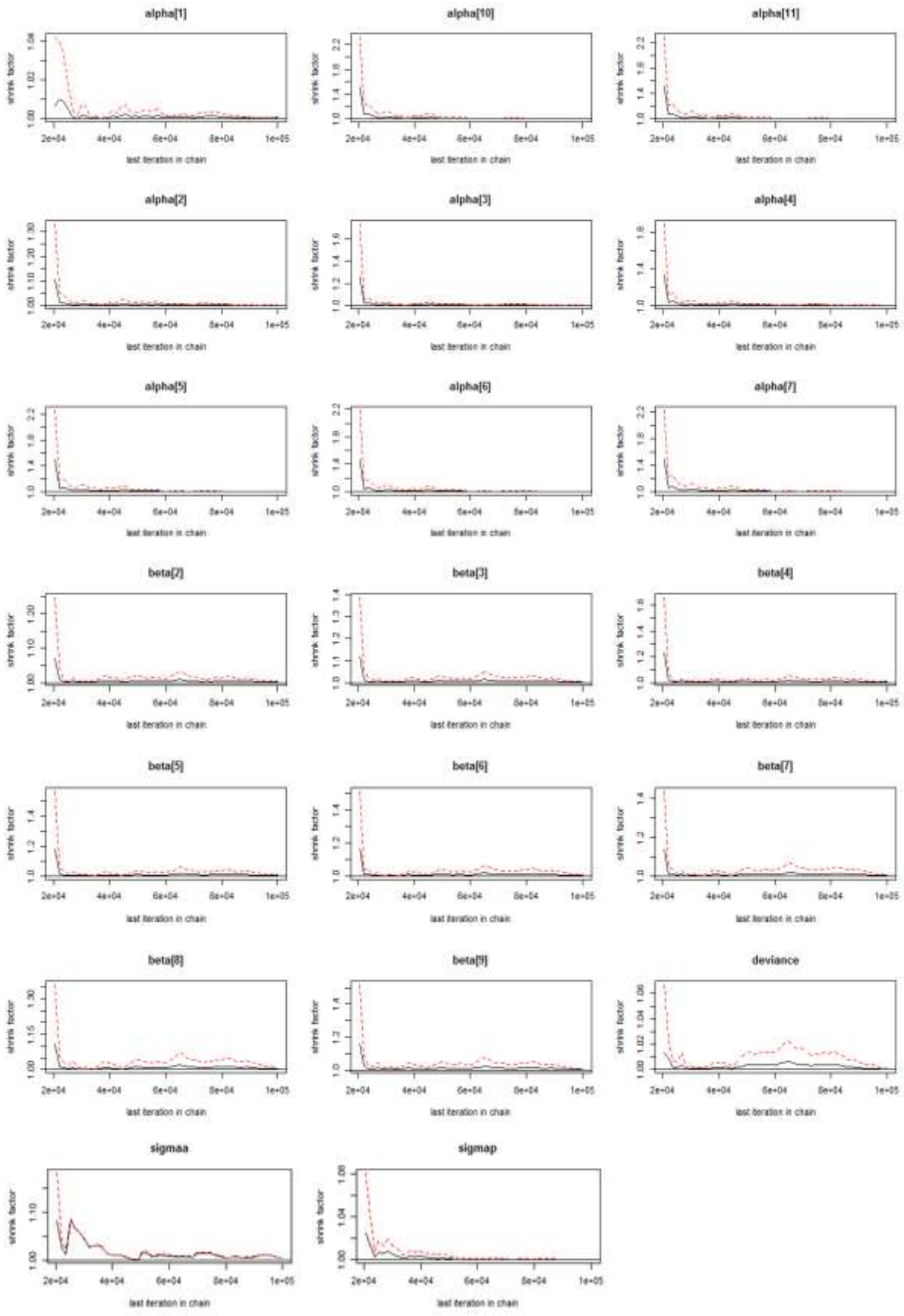


Figure 7.5. Plots of Gelman-Rubin's diagnostic of selected parameters of the AP model.

7.6. Prediction to 2020

This section presents the projections from 2010 to 2020 using our chosen best model. Here we observed that the AP model performed better than the APC and AC models (see Table 7.14). The posterior mean and the 95% credible intervals were obtained after the last 100,000 iterations. Fitted and projected rates were then estimated from samples of 100,000 drawn from the posterior distribution after excluding the first 20,000 iterations as burn-in. We present the graphs in Figure 7.6 below.

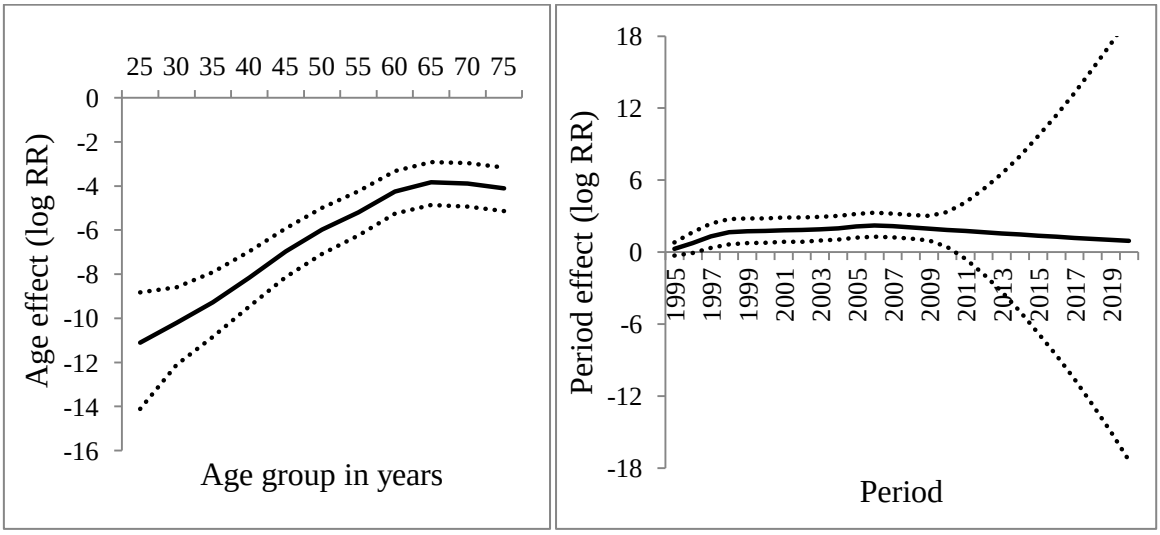


Figure 7.6: Age and period effects, on lung cancer mortality in KSA identified by AP model from age 25 to 75 and over during the period 1994-2020 within 95% credible intervals (dash lines.....).

Figure 7.6 above shows an increase of lung cancer mortality up to the age class 65-69 and then it starts to decrease gradually for the rest of age groups whereas, the period effects declined slightly over time. Between 1998 and 2009 the period effect on lung cancer mortality rate has increased dramatically by 0.2% every year. However, our forecasts show that between 2009 and 2020, lung cancer period effect is expected to decrease by 0.5% per year. Note that the projected period effect is uncertain. The credible intervals include both uncertainty associated with the choice of the model and uncertainty associated with forecasting beyond the range of the data. This is necessarily reflected by rapidly increasing width of intervals as the length of projection increases.

Table 7.15: The effects of age and period on lung cancer mortality in KSA estimated from the Bayesian dynamic AP model from 1994 to 2020.

Age effect ($\hat{\alpha}_i$)	Estimates	Standard Errors	95 % Credible Intervals
25-29	-11.11	1.35	-14.11, -8.83
30-34	-10.20	0.89	-12.10, -8.60
35-39	-9.27	0.74	-10.84, -7.91
40-44	-8.16	0.64	-9.47, -6.96
45-49	-6.97	0.56	-8.13, -5.92
50-54	-5.98	0.53	-7.08, -4.99
55-59	-5.20	0.51	-6.24, -4.24
60-64	-4.24	0.49	-5.26, -3.32
65-69	-3.84	0.50	-4.86, -2.92
70-74	-3.89	0.50	-4.93, -2.95
75+	-4.10	0.50	-5.13, -3.17
Period effect ($\hat{\beta}_i$)			
1995	0.26	0.28	-0.28, 0.82
1996	0.76	0.44	-0.07, 1.66
1997	1.32	0.52	0.34, 2.38
1998	1.65	0.54	0.65, 2.75
1999	1.74	0.53	0.75, 2.81
2000	1.77	0.52	0.79, 2.82
2001	1.82	0.52	0.86, 2.88
2002	1.84	0.51	0.87, 2.88
2003	1.91	0.51	0.96, 2.95
2004	2.00	0.51	1.04, 3.02
2005	2.15	0.50	1.20, 3.18
2006	2.24	0.51	1.30, 3.28
2007	2.18	0.50	1.24, 3.22
2008	2.06	0.50	1.12, 3.10
2009	1.96	0.52	0.97, 3.03
2010	1.86	0.75	0.42, 3.36
2011	1.76	1.20	-0.56, 4.15
2012	1.66	1.80	-1.83, 5.25
2013	1.57	2.50	-3.31, 6.54
2014	1.47	3.29	-4.96, 8.00
2015	1.37	4.16	-6.74, 9.66
2016	1.28	5.09	-8.59, 11.41
2017	1.19	6.09	-10.57, 13.26
2018	1.10	7.15	-12.75, 15.29
2019	1.02	8.27	-14.99, 17.38
2020	0.93	9.44	-17.36, 19.58

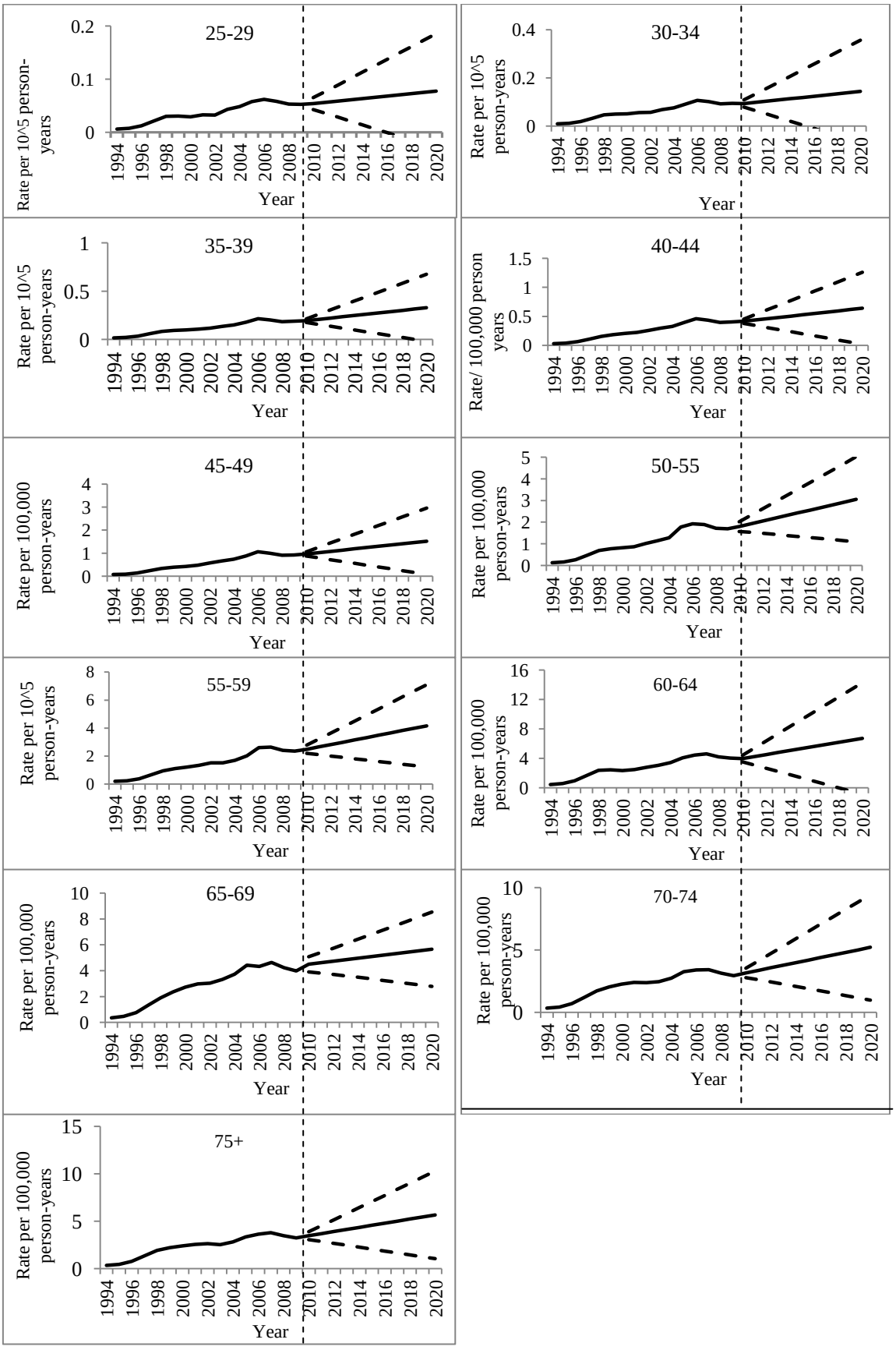


Figure 7.7: Fitted (1994-2009) and projected (2010-2020) age-specific standardized rate of lung cancer mortality (per 100,000 person-year) in Saudi Arabia, with 95% credible intervals (dashed lines---), for each 5 year age-group in the range 25-75 years based on the final Bayesian AP model.

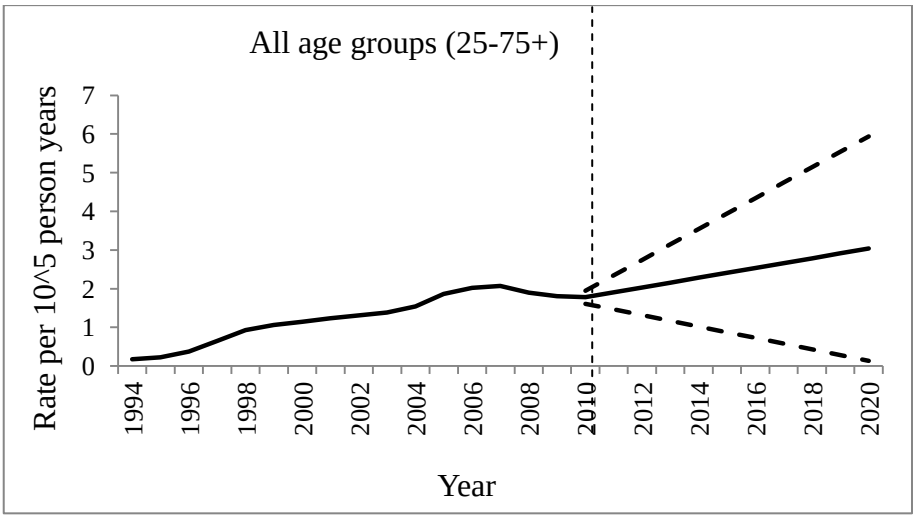


Figure 7.8: Fitted and projected age standardized rate of lung cancer mortality (per 100,000 person-year) in Saudi Arabia for age groups 25-75 years up to 2020, according to the final Bayesian AP model with 95% credible intervals for the projection (dashed lines ----).

7.7. Discussions

Separating the effects of age, period and cohort is challenging because of the identifiability of the parameters. However, in the last thirty years several methods have been suggested to overcome this identification problem e.g. spline functions and Bayesian dynamic APC. In this chapter, we use Bayesian dynamic age-period-cohort models to solve the identification problem because the number of cases of the lung cancer mortality is low. In this Bayesian analysis, trends were modelled through specific smoothing of model parameters using RW2 for all the three effects of age, period and cohort. Results from Bayesian model with reduced parameters of age and period effects are almost identical to those from age, period and cohort effects, suggesting that Bayesian AP model is preferred, in line with the recommendation of Clayton and Schifflers (1987a and 1987b). They advised to reduce the APC model to either an age-period (AP) model or an age-cohort (AC) model, whichever is better, and to only use the APC model when it provides a more satisfactory fit. The prior distribution used in our analysis is a non-informative uniform distribution because we want the hyper-parameters to be estimated mainly from the data.

Our results show that the most important effect on lung cancer is age followed by the period. The age effect shows a dramatic increase of lung cancer mortality up to the age class 65-69 and then it starts to decrease slightly for the rest of age groups. The Bayesian dynamic age-period model shows that the effect of the period reached its maximum in 2007, as an effect of the new polices implemented by the government during the period, such as the ban of smoking cigarettes in public places and the increase of tax on imported

tobacco. The cohort effect is not important due to the short time period of the observed mortality data and therefore is assumed to be equal over different generations.

Our results suggests that the expected age-specific standardized rates of lung cancer mortality will increase gradually in all age groups between 2010 and 2020. For instance, in the age group 50-54 the posterior mean (age-specific standardized rate) within its 95% credible interval is expected to increase from 1.8 (1.57, 2.02) in 2010 to 3.06 (1.09, 5.02) per 100,000 population in 2020 whereas in the age group 65-69 is expected to increase from 4.50 (3.90, 5.09) in 2010 to 5.66 (2.78, 8.54) per 100,000 population in 2020. Overall, the risk of mortality reaches its peak from lung cancer in Saudi Arabia between ages 60 and 70.

Our Bayesian dynamic AP model provides a good fit to the mortality of lung cancer rates compared to the Bayesian dynamic AC and APC models. The fitted rate from age effect shows that the mortality rate increases as the age increases with wider credible interval width at younger age groups. The width of the intervals is due to the heterogeneity in the data where there are sparse, zero counts and uncertainty associated with the fitted model. This can be seen by the sudden increase of lung cancer mortality rate in the age group between the range 25-49 years due to the sparse and zero counts. This is shown in the mortality data in Table F15 and F16 in Appendix F. For the period effect, there is a gradual decrease in the rate-ratio over time. However, the projected period effect is uncertain. The credible intervals include both uncertainty associated with the choice of the model and uncertainty associated with forecasting beyond the range of the data. This is necessarily shown by rapidly increasing width of intervals as the length of prediction increases. The age effect is much stronger than the period effect as shown in Figure 7.6 and Table 7.15 above. The 95% credible interval is much tighter in the age effect than in the period effect so the credible interval is dominated by the uncertainty in the age effect rather than the uncertainty in the period effect.

7.8. Summary

The observed data used for Bayesian dynamic age-period (AP) modelling were annual mortality cases of lung cancer, for Saudi Arabia, by combined gender and 5-year age group in the range 25-75 years between 1994 and 2009. The projection of the future rate depends on population projections; we use the 2010 to 2020 United Nations forecasts for KSA. These projections do not take into account potential changes in lifestyle or treatment that could alter future rates of lung cancer mortality.

Overall, the estimated age standardized rate (ASR) of lung cancer mortality from the Bayesian dynamic AP model shows a gradual increase between 1994 and 2007. However, it decreased in 2008 and started to increase again in 2009. Our projection shows that the ASR of lung cancer mortality is expected to increase to 2020, from 1.8 (1.61, 1.94) in 2010 to 3.04 (0.13, 5.94) per 100,000 population in 2020. The trends are mainly due to the age effect and slightly due to the period effect but no obvious cohort effects were observed in our study. The lack of cohort effect may be due to the short time period of the observed mortality data. Age has a strong association with lung cancer mortality, suggesting age-related causes such as accumulative exposures of smoking over time may be the main reason for increasing lung cancer mortality in KSA, since the prevalence of smoking is increasing especially in women. Tobacco is responsible for around 70% of lung cancer mortality (World Health Organization, Media Center, 2015). The increase of lung cancer mortality rate in all age groups during the period of our study could also be attributed to the lack of early detection and screening of lung cancer mortality.

In summary, the increased lung cancer mortality rate from 1994 to 2009 is mostly attributed to age effects. The ASR of lung cancer mortality will increase gradually until 2020. Lung cancer burden will continue to increase due to the aging population and may be due to the increase of smoking prevalence especially in women.

Notice that in this analysis we did not include any covariate variables because our main aim here is to produce quite reliable estimates of future lung cancer mortality in KSA since our lung cancer mortality data is low. Additionally, the use of APC models are identified as proxies for events such as risk factors, which we cannot measure directly.

The lack of sufficient data is due to the late establishment of the Saudi Cancer Registry (SCR) in 1992. However, the data included in this study were compiled only up to December 2009 because it takes about 3-4 years before the data is processed for public consumption. In addition, the SCR has to wait until all the data from the regional cancer registries have been collected. Notwithstanding, there are a number of complex processes to register a cancer case in order to ensure the data is of a high quality (Al-Eid, Saudi Cancer Incidence Report, 2009). Therefore, the Ministry of Health should pay attention to reduce the time length of collecting the data from the branches as much as possible in order to improve short-term forecasts. Thus, we suggest that the Government should establish additional main cancer registries in Western, Eastern, Northern and Southern regions because of the geographical and demographical characteristics of the country with the required implementation materials and training methods needed for the staff.

CHAPTER 8

CONCLUSIONS AND RECOMMENDATIONS

8.1. Conclusions

Cancer is a global health challenge. It is perhaps the most significant problem which humanity will have to face in the next two or more decades after global warning (World Health Organization, Media Center, 2015). Nowadays lung cancer is the first or second most frequent tumor type among men and third or fourth among women (World Health Organization, Media Center, 2015). Therefore, efforts to reduce and prevent lung cancer are of course essential.

Forecasting the burden of lung cancer incidence and mortality is important for evaluating prevention strategies and for administrative planning at lung cancer facilities. We collect the data of lung cancer incidence and mortality from Saudi Cancer Registry (SCR) and Central Department of Statistics (CDS) in Saudi Arabia (KSA) from 1994 to 2009. Population data were prepared from forecasts made by the United Nations between 2010 and 2020. Our aim is to use forecasting methods to describe the broad picture of the future lung cancer burden in Saudi Arabia and to report baseline incidences against which progress in implementing the National Health Service (NHS) Cancer Plan will be measured.

We study lung cancer incidence and mortality in Saudi Arabia between 1994 and 2009. The first part of this study uses time-series methods in determining and forecasting lung cancer incidence data using Box-Jenkins methodology and dynamic regression models. In the Box-Jenkins analysis, we present Seasonal Autoregressive Integrated Moving Average (SARIMA) models in chapter 4. In dynamic regression, we describe more general autoregressive AR processes such as AR(1), distributed lag models (DLMs), and polynomial distributed lag models (PDLs). We develop, analyze, and perform a one-step ahead forecast of the various models to explore the best-fit model for lung cancer cases in Saudi Arabia. We propose a new approach called autoregressive polynomial distributed lag (ARPDL) model. This approach results in having a model with a lower standard error and more accurate fit. The second part of this study concentrates on the age-period-cohort (APC) models. Natural cubic splines were used in APC models for drawing inference on the impact of lung cancer incidence rates. Using the restriction of the cubic splines being linear beyond the boundary knots, we were able to make better projections in

the magnitude of the rates, the variation by age, and time trends in the rates into the future. Using splines and more finely split data as opposed to the factor models with coarsely split data seems better. Bayesian dynamic APC models were used for modelling and forecasting lung cancer mortality rates between 1994 and 2020. Bayesian approaches assume some sort of smoothness of age, period and cohort effects in order to improve estimation and facilitate prediction. Three models were used: the full APC, AP and AC models. Comparison between nested models was evaluated by the changes in Deviance Information Criterion DIC.

The empirical results of lung cancer incidence show that most of the cases are among males and suggest that lung cancer cases are strongly affected by smoking habits. The overall best one-step-ahead forecast of dynamic regression model is the ARPDL(12,3,26,8) model of the total cases of lung cancer on smoking population separately for males and females. This is confirmed by the value of adjusted R-squared as well as the significance of the F-statistic of the regressions. The overall best Box-Jenkins SARIMA model is the SARIMA(2,1,1)x$(0,1,1)_{12}$ model. It is best on all three information criteria: AIC, AICc and BIC. The forecasts generated by ARPDL and SARIMA models both capture the seasonality trends. However, the ARPDL model with a small lag does not capture the seasonality as well as the ARPDL model with large lag. Nonetheless, we prefer the forecast generated from the SARIMA model since it has a fewer parameters to estimate. The preferred SARIMA model suggests that there will be an average of 45 cases of lung cancer per month for the next 24 months. In addition, the estimated yearly lung cancer cases in 2010 and 2011 were 538 and 555 respectively. We conclude from the data that more incident cases are diagnosed in winter.

The estimated incidence rates from age-period-cohort modelling show a sharp decrease in males and a gradual increase in females over the next 10 years. The male age standardised rate of lung cancer incidence is projected to fall from 5.6 to 2.4 per 100,000 by 2020, whereas the female age standardised rate of lung cancer incidence is projected to increase from 2.0 to 2.2 per 100,000 by 2020. The growing and ageing populations will have a substantial impact, therefore the number of cases per year are projected to decrease in males (from 356 to 320) and to increase in females (from 134 to 247) between 2009 and 2020. These results reflect the decrease of smoking prevalence among males and the increase of smoking prevalence in females. The results show that in KSA, males have about a 79% greater incidence of lung cancer than females across the entire dataset when adjusting for the other effects. The p-value for the gender term highlights that the effect for

gender is significant at the 0.1% level. In addition, the p-values for the covariates of race, Southern, Western, and Eastern regions show that the effects for these covariates are statistically significant.

By comparing the trends of lung cancer incidence in Saudi Arabia (KSA) to that of the United Kingdom (UK), we seem to have almost the same pattern. However, the rate of lung cancer incidence is much higher in the UK than in KSA due to the high prevalence of smoking among males and females in the UK (see Appendix B2 and B3). In 1994, the overall age-standardised incidence rates of lung cancer in the UK were 90.5 and 35 per 100,000 for males and females respectively. Over the same period, the overall age-standardised incidence rates in KSA were 7.7 and 2 per 100,000 for males and females respectively. The projection of lung cancer incidence cases from 2009 to 2020 for both countries is expected to decrease sharply in males by 16.28% in UK and 57.14% in KSA. On the other hand, females are expected to have a slight decrease by 8.45% in UK and a slight increase by 10% in KSA. Thus, age-standardised incidence rates are projected to decease in males to 47.8 and to 2.4 per 100,000 in the UK and in KSA respectively. Whereas females age-standardised incidence rates are expected to decrease slightly in the UK to 32.5 and to increase slightly in KSA to 2.2 per 100,000.

The estimated age standardized rate (ASR) of lung cancer mortality within its 95% credible interval is expected to increase from 1.8 (1.61, 1.94) in 2010 to 3.04 (2.13, 5.94) per 100,000 population in 2020. Our results suggest that the expected age-specific standardized rates of lung cancer mortality will increase gradually in all age groups between 2010 and 2020. Mortality risk from lung cancer reaches its peak between ages 65 and 69 years. The posterior mean (age-specific standardized rate) within its 95% credible interval is expected to increase from 4.50 (3.90, 5.09) in 2010 to 5.66 (2.78, 8.54) per 100,000 population in 2020. The trends of lung cancer mortality are mainly due to the age effect and slightly due to the period effect but no obvious cohort effects were observed in the study. The lack of cohort effect may be due to the short time period of the observed mortality data. Age has a strong association with lung cancer mortality, suggesting age-related causes such as cumulative exposures of smoking over time. This may be the main reason of increasing lung cancer mortality in KSA, since the prevalence of smoking is increasing especially in women. Tobacco is responsible for around 70% of lung cancer mortality (World Health Organization, Media Center, 2015). The increase of lung cancer mortality rate in all age groups during the period of our study could also be attributed to the lack of early detection and screening of lung cancer mortality.

In this thesis we have proposed different approaches to model and forecast lung cancer incidence and mortality in Saudi Arabia. We used finite and infinite dynamic regression models and we came up with a new approach called autoregressive polynomial distributed lag (ARPDL) model. This approach results in having a model with a lower standard error and more accurate fit than PDL and OLS models. Also, we used two methodological approaches on modelling age-period-cohort models, namely spline functions and Bayesian dynamic models. Our results show that both APC models using spline functions and Bayesian dynamic models are able to overcome the identification problem and identify the effect of age, period and cohort. However, Bayesian dynamic APC model is preferred in forecasting the incidence or the mortality rates of lung cancer especially when the data are sparse or has zero counts, because the forecast based on Bayesian dynamic APC model does not rely on strong parametric assumptions for future values of subjective cohort and period effects. Additionally, the sparse data and zero counts in Bayesian dynamic APC models do not pose any implementation problems when fitting APC models.

8.2. Limitations of the Work

It is important to recognize and highlight potential limitations in our data and methods to ensure that results and findings obtained are reliable. Data quality can be an issue, as it will have implications for the confidence that can be placed in a study output. Although the data size is not large, checking the residuals for normality each time after performing normal regression on a model shows that the residuals are normally distributed, indicating minimal noise in the data. Another limitation is that, historical data are often assumed to be correct, without any means of assessing whether or not they were collected, processed or interpreted adequately. There is mis-recording as some cohort date of birth, age, dead or alive values are recorded as zero.

The use of SARIMA models has some limitations: first, some of the traditional model identification techniques for identifying the correct model order from the class of possible models are not clear cut. This process is also subjective and the reliability of the chosen model can depend on the skill and experience of the researcher. Second, the underlying theoretical model and structural relationships are difficult to apply (O‟Donovan, 1983).

The simplest way to estimate parameters associated with distributed lags (unrestricted) is by ordinary least squares. However, multicollinearity among the lagged explanatory variables often arises, leading to large variance of the coefficient estimates. There are two disadvantages to the finite distributed lag model. The first is multicollinearity. The second

disadvantage of finite distributed lags is that they can be problematic when the lag length is large, especially in small samples. Estimation of the infinite distributed lag (Koyck) model also presents some challenges because the lagged independent variable is by definition not strictly exogenous and, unless the error term is white noise, is not even weakly exogenous.

In summary, the finite distributed lag (Almon) model is most suitable to estimating dynamic relationships when lag weights decline to zero relatively quickly, when the regressor is not highly autocorrelated, and when the sample is long relative to the length of the lag distribution. However, the finite distributed lag models are not without problems. The polynomial distributed lag model allows the data to determine the shape of the lag structure, but the researcher must choose the maximum lag length, choose the degree of the polynomial, and overcome the difficulty in capturing long-tailed lag distributions. An incorrectly specified maximum lag length and the rest can distort the shape of the estimated lag structure as well as the cumulative effect of the independent variable.

The impossibility to attribute the drift to respectively cohort or period related effects, because of their linear dependency, implies a serious problem in displaying and estimating the model parameters (Clayton and Schifflers, 1987). Nevertheless, APC-modelling protects against over-interpretation of trends based on standardised rates or simple graphical presentation of age-specific curves.

The use of Bayesian dynamic APC models also have some limitations. Firstly, the projected period effect is uncertain. The credible intervals include both uncertainty associated with the choice of the model and uncertainty associated with forecasting beyond the range of the data. This is necessarily reflected by rapidly increasing width of intervals as the length of projection increases. Secondly, these projections do not take into account potential changes in lifestyle or treatment that could alter future rates of lung cancer mortality. Thirdly, although our models did not show any convergence problems with the use of MCMC, the use of MCMC algorithms with random walk models has been criticized in terms of both computational time and mixing due to strong dependencies of parameters in the posterior distribution and of week identifiability. Thus, integrated nested Laplace approximations should be used to overcome these problems as an alternative solution (Carreras and Gorini, 2014).

8.3. Recommendations

The government should monitor the rise of lung cancer cases during winter to consider providing more health-care resources in winter. Government should make it a priority in its policy agenda during this period to provide more training for health staff in various treatments.

The Ministry of Health should investigate and plan new strategies in areas where they are affected such as Southern, Western, and Eastern regional hospitals. . The Government should discourage tobacco advertising, promotion, and sponsorship in order to reduce the prevalence of smoking as much as possible since tobacco is responsible for around 70% of lung cancer mortality (World Health Organization, Media Center, 2015).

The Ministry of Health should pay attention to reduce the time length of collecting the data from the branches as much as possible in order to improve short-term forecasts. Thus, we suggest that the government should establish additional main cancer registries in Western, Eastern, Northern and Southern regions because of the geographical and demographical characteristics of the country with the required implementation materials and training methods needed for the staff.

8.4. Future Research

Having determined the best models used in forecasting lung cancer incidence and mortality rates, the next stage of our research will focus on forecasting breast cancer in KSA. To do this, we describe a detailed plan next.

A dynamic Poisson model will be used with a Bayesian approach to modelling to predict breast cancer incidence and mortality in Saudi Arabia. The complexity of the posterior distribution prohibits direct evaluation of the posterior, and therefore parameters will be estimated by the recently proposed Integrated Nested Laplace Approximations (INLA). INLA is a promising alternative to inference via MCMC in latent Gaussian models (Rue et al., 2009). INLA is a useful and flexible tool for Bayesian hierarchical models with complex dependence structure with loads of linear constraints. The out-of-sample forecast is straightforward and the running time is fast (Held, 2009).

We will continue to work on extending our methodology to a more general Autoregressive Distributed Lag ARDL(p,q) models and a more general Moving Average Distributed Lag MADL(p,q) models. These will include both incidence and mortality. We will also explore the possibilities of addressing additional covariates using APC models.

APPENDICES

Appendix A: Results of Dynamic Regression Models.

Table A1: Estimated slope from OLS regression through the origin of residuals on lagged residuals.

```
The regression equation is
RESI1 = 0.173 lagged res
191 cases used, 1 cases contain missing values

Predictor      Coef  SE Coef     T      P
Noconstant
lagged res  0.17259  0.06787  2.54  0.012

S = 6.64865
Analysis of Variance
Source           DF       SS      MS     F      P
Regression        1   285.90  285.90  6.47  0.012
Residual Error  190  8398.87   44.20
Total           191  8684.77
```

Table A2: Estimated rho from Cochrane-Orcutt iterative procedure.

```
. prais Yt Xt, corc

Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.1726
Iteration 2:  rho = 0.1726
Iteration 3:  rho = 0.1726

Cochrane-Orcutt AR(1) regression -- iterated estimates
      Source |       SS       df       MS              Number of obs =     191
-------------+------------------------------           F(  1,   189) =  111.26
       Model |  4938.19512     1  4938.19512           Prob > F      =  0.0000
    Residual |   8388.7016   189  44.3846646           R-squared     =  0.3705
-------------+------------------------------           Adj R-squared =  0.3672
       Total |  13326.8967   190  70.1415617           Root MSE      =  6.6622
------------------------------------------------------------------------------
          Yt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          Xt |    .120256   .0114009    10.55   0.000     .0977666    .1427454
       _cons |  -1.344754   3.232239    -0.42   0.678    -7.720653    5.031146
-------------+----------------------------------------------------------------
         rho |   .1726466
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    1.555379
Durbin-Watson statistic (transformed) 2.002804
```

Table A3: Results of Prais-Winsten iterative procedure.

```
.  prais Yt Xt

Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.1726
Iteration 2:  rho = 0.1728
Iteration 3:  rho = 0.1728
Prais-Winsten AR(1) regression -- iterated estimates

      Source |       SS       df       MS              Number of obs =     192
-------------+------------------------------           F(  1,   190) =   99.62
       Model |  4881.25627     1  4881.25627           Prob > F      =  0.0000
    Residual |  9310.16083   190  49.0008465           R-squared     =  0.3440
-------------+------------------------------           Adj R-squared =  0.3405
       Total |  14191.4171   191  74.3006131           Root MSE      =  7.0001

------------------------------------------------------------------------------
          Yt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          Xt |   .1157449   .0119365     9.70   0.000     .0921998      .13929
       _cons |   .1402733    3.37961     0.04   0.967    -6.526104     6.80665
-------------+----------------------------------------------------------------
         rho |   .1728324
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    1.555379
Durbin-Watson statistic (transformed) 1.935193
```

Table A4: Results of Hildreth-Lu search procedure.

```
. hlu Yt Xt

Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.9999
Iteration 2:  rho = 0.5000
Iteration 3:  rho = 0.2500
Iteration 4:  rho = 0.3750
Iteration 5:  rho = 0.3125
Iteration 6:  rho = 0.2812
Iteration 7:  rho = 0.2656
Iteration 8:  rho = 0.2578
Iteration 9:  rho = 0.2539
Iteration 10: rho = 0.2519

(Hildreth-Lu regression)
      Source |       SS       df       MS              Number of obs =     191
-------------+------------------------------           F(  1,   189) =   91.25
       Model |   4079.1431     1   4079.1431           Prob > F      =  0.0000
    Residual |  8449.03314   189   44.703879           R-squared     =  0.3256
-------------+------------------------------           Adj R-squared =  0.3220
       Total |  12528.1762   190  65.9377697           Root MSE      =  6.6861

------------------------------------------------------------------------------
          Yt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          Xt |     .12076   .0126419     9.55   0.000     .0958235    .1456964
      _inter |  -1.502849   3.585433    -0.42   0.676    -8.575218    5.569519
-------------+----------------------------------------------------------------
         rho |     0.2510     0.0678     3.70   0.000       0.1173      0.3846
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    1.555379
Durbin-Watson statistic (transformed) 2.170192
```

Table A5: Regression results of total cases of lung cancer against smoking population

```
The regression equation is
Yt = - 0.03 + 0.116 Xt

Predictor     Coef  SE Coef      T      P
Constant    -0.034    2.849  -0.01  0.990
Xt         0.11623  0.01007  11.55  0.000

S = 7.11722   R-Sq = 41.2%   R-Sq(adj) = 40.9%
Analysis of Variance
Source           DF       SS      MS       F      P
Regression        1   6753.9  6753.9  133.33  0.000
Residual Error  190   9624.4    50.7
Total           191  16378.3
Durbin-Watson statistic = 1.55538
```

Table A6: Regression results of total cases of lung cancer against smoking population with lag one

```
The regression equation is
Yt = - 2.74 - 1.11 Xt + 1.24 Xt-1

191 cases used, 1 cases contain missing values
Predictor     Coef  SE Coef      T      P
Constant    -2.735    2.696  -1.01  0.312
Xt         -1.1126   0.3659  -3.04  0.003
Xt-1        1.2414   0.3686   3.37  0.001
S = 6.59673   R-Sq = 48.1%   R-Sq(adj) = 47.6%

Analysis of Variance
Source           DF       SS      MS      F      P
Regression        2   7585.3  3792.6  87.15  0.000
Residual Error  188   8181.2    43.5
Total           190  15766.4
Durbin-Watson statistic = 1.7100
```

Table A7: Regression results of total cases of lung cancer against smoking population with six lags

```
The regression equation is

Yt = - 4.24 - 0.464 Xt - 0.18 Xt-1 + 1.39 Xt-2 - 1.12 Xt-3 + 0.47 Xt-4
     - 0.18 Xt-5 + 0.219 Xt-6

186 cases used, 6 cases contain missing values

Predictor     Coef  SE Coef      T      P
Constant    -4.244    2.881  -1.47  0.142
Xt         -0.4641   0.7901  -0.59  0.558
Xt-1        -0.176    1.715  -0.10  0.918
Xt-2         1.391    1.864   0.75  0.457
Xt-3        -1.120    1.864  -0.60  0.549
Xt-4         0.470    1.864   0.25  0.801
Xt-5        -0.185    1.719  -0.11  0.914
Xt-6        0.2191   0.8008   0.27  0.785
S = 6.67538   R-Sq = 49.3%   R-Sq(adj) = 47.4%
Analysis of Variance
Source           DF       SS      MS      F      P
Regression        7   7727.3  1103.9  24.77  0.000
Residual Error  178   7931.8    44.6
Total           185  15659.1

Durbin-Watson statistic = 1.72907
```

Table A8: Results of Koyck transformation.

```
The regression equation is

Yt = - 1.18 + 0.0994 Xt + 0.176 Yt-1
191 cases used, 1 cases contain missing values

Predictor      Coef  SE Coef      T      P
Constant     -1.182    2.682  -0.44  0.660
Xt          0.09939  0.01221   8.14  0.000
Yt-1        0.17621  0.06786   2.60  0.010
S = 6.67419   R-Sq = 46.9%   R-Sq(adj) = 46.3%
Analysis of Variance
Source           DF       SS      MS      F      P
Regression        2   7392.0  3696.0  82.97  0.000
Residual Error  188   8374.4    44.5
Total           190  15766.4

Durbin-Watson statistic = 2.01369
```

Table A9: Minitab output for AR(1).

```
Final Estimates of Parameters

Type          Coef  SE Coef      T      P
AR   1      0.5462   0.0613   8.91  0.000
Constant  14.7881   0.5650  26.17  0.000
Mean       32.584    1.245
Number of observations:  192

Residuals:    SS =  11634.2 (backforecasts excluded)
              MS =  61.2  DF = 190
Modified Box-Pierce (Ljung-Box) Chi-Square statistic
Lag             12     24     36     48
Chi-Square    62.3  108.8  138.0  158.6
DF              10     22     34     46
P-Value      0.000  0.000  0.000  0.000

Forecasts from period 192
                     95 Percent
                       Limits
Period  Forecast    Lower    Upper  Actual
   193   41.5495  26.2091  56.8898
```

Table A10: Regression results.

```
The regression equation is

Yt = - 1.447 + 0.1211 Xt-1

S = 6.73908   R-Sq = 45.6%   R-Sq(adj) = 45.3%

Analysis of Variance

Source       DF       SS       MS       F      P
Regression    1   7183.0  7182.97  158.16  0.000
Error       189   8583.5    45.42
Total       190  15766.4
```

Table A11: Results of Cochrane-Orcutt AR (1) iterative procedure.

```
. prais Yt Xt1, corc

Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.1806
Iteration 2:  rho = 0.1806
Iteration 3:  rho = 0.1806
Cochrane-Orcutt AR(1) regression -- iterated estimates

      Source |       SS       df       MS              Number of obs =     190
-------------+------------------------------           F(  1,   188) =  109.59
       Model |  4840.45371     1  4840.45371           Prob > F      =  0.0000
    Residual |  8303.35851   188  44.1668006           R-squared     =  0.3683
-------------+------------------------------           Adj R-squared =  0.3649
       Total |  13143.8122   189    69.54398           Root MSE      =  6.6458

------------------------------------------------------------------------------
          Yt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         Xt1 |   .1213988   .0115963    10.47   0.000     .0985233    .1442744
       _cons |  -1.539908   3.281055    -0.47   0.639    -8.012324    4.932507
-------------+----------------------------------------------------------------
         rho |    .180584
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    1.637079
Durbin-Watson statistic (transformed) 2.037903
```

Table A12: Results of DLM using Koyck transformation.

```
The regression equation is
Yt = - 1.49 + 0.102 Xt-1 + 0.169 Yt-1


191 cases used, 1 cases contain missing values

Predictor     Coef  SE Coef      T      P
Constant    -1.486    2.683  -0.55  0.580
Xt-1       0.10161  0.01230   8.26  0.000
Yt-1       0.16890  0.06786   2.49  0.014


S = 6.64834   R-Sq = 47.3%   R-Sq(adj) = 46.7%

Analysis of Variance

Source           DF       SS      MS      F      P
Regression        2   7456.8  3728.4  84.35  0.000
Residual Error  188   8309.7    44.2
Total           190  15766.4
```

Table A13: Results of Breusch-Godfrey LM test of ARPDL(12,5,26,8) model.

F-statistic	1.277192	Prob. F(1,149)	0.2602
Obs*R-squared	1.410819	Prob. Chi-Square(1)	0.2349

Test Equation:
Dependent Variable: RESID
Method: Least Squares
Date: 12/21/14 Time: 19:40
Sample: 1996M03 2009M12
Included observations: 166
Presample missing value lagged residuals set to zero.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.089198	4.907420	-0.629495	0.5300
PDL01	0.011330	0.137187	0.082586	0.9343
PDL02	-0.036211	0.083805	-0.432084	0.6663
PDL03	0.001062	0.023269	0.045621	0.9637
PDL04	0.001835	0.004125	0.444840	0.6571
PDL05	-7.97E-05	0.000805	-0.098927	0.9213
PDL06	-2.36E-05	5.47E-05	-0.432072	0.6663
PDL07	1.22E-06	9.80E-06	0.124544	0.9011
PDL08	8.82E-08	2.19E-07	0.401715	0.6885
PDL09	-5.12E-09	3.69E-08	-0.138697	0.8899
PDL010	-0.069043	0.075845	-0.910318	0.3641
PDL011	-0.006449	0.028601	-0.225478	0.8219
PDL012	0.015732	0.016660	0.944298	0.3465
PDL013	-0.001349	0.003779	-0.357054	0.7216
PDL014	-0.000789	0.000785	-1.005058	0.3165
PDL015	0.000113	0.000140	0.804296	0.4225
RESID(-1)	-0.175670	0.305850	-0.574366	0.5666

R-squared	0.008499	Mean dependent var	1.01E-11
Adjusted R-squared	-0.097971	S.D. dependent var	5.988542
S.E. of regression	6.275041	Akaike info criterion	6.607815
Sum squared resid	5867.044	Schwarz criterion	6.926512
Log likelihood	-531.4486	Hannan-Quinn criter.	6.737176
F-statistic	0.079824	Durbin-Watson stat	1.996212
Prob(F-statistic)	0.999999		

Table A14: Results of Breusch-Godfrey LM test of ARPDL(12,3,26,8) model.

F-statistic	0.090604	Prob. F(1,142)	0.7639
Obs*R-squared	0.105850	Prob. Chi-Square(1)	0.7449

Test Equation:
Dependent Variable: RESID
Method: Least Squares
Date: 01/07/15 Time: 22:18
Sample: 1996M03 2009M12
Included observations: 166
Presample missing value lagged residuals set to zero.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.433135	18.82833	-0.076116	0.9394
PDL01	0.003477	0.165845	0.020963	0.9833
PDL02	0.011342	0.104549	0.108483	0.9138
PDL03	-0.000755	0.027244	-0.027703	0.9779
PDL04	-0.000452	0.004886	-0.092434	0.9265
PDL05	2.32E-05	0.000937	0.024723	0.9803
PDL06	4.55E-06	6.30E-05	0.072220	0.9425
PDL07	-2.54E-07	1.15E-05	-0.022016	0.9825
PDL08	-1.35E-08	2.52E-07	-0.053338	0.9575
PDL09	9.03E-10	4.38E-08	0.020588	0.9836
PDL010	-0.038243	0.494265	-0.077373	0.9384
PDL011	-0.013540	0.278196	-0.048669	0.9613
PDL012	0.004063	0.083870	0.048446	0.9614
PDL013	0.000220	0.013950	0.015774	0.9874
PDL014	-5.59E-05	0.002878	-0.019436	0.9845
PDL015	2.01E-06	0.000194	0.010386	0.9917
PDL016	-1.11E-08	3.46E-05	-0.000321	0.9997
PDL017	-2.18E-08	8.09E-07	-0.026963	0.9785
PDL018	1.56E-09	1.29E-07	0.012045	0.9904
PDL019	0.005785	0.050490	0.114571	0.9089
PDL020	0.002095	0.014163	0.147903	0.8826
PDL021	0.000416	0.002214	0.187924	0.8512
PDL022	-0.000187	0.000907	-0.206658	0.8366
RESID(-1)	-0.075617	0.145462	-0.519842	0.6040

R-squared	0.000638	Mean dependent var	1.38E-11
Adjusted R-squared	-0.161231	S.D. dependent var	5.356233
S.E. of regression	5.771900	Akaike info criterion	6.476876
Sum squared resid	4730.705	Schwarz criterion	6.926801
Log likelihood	-513.5807	Hannan-Quinn criter.	6.659503
F-statistic	0.003939	Durbin-Watson stat	2.000741
Prob(F-statistic)	1.000000		

Appendix B: Cancer Incidence Forecast in UK up to 2020

Appendix B1: Different Cancer Incidence Rates Diagnosed in UK from 1980 to 2030.

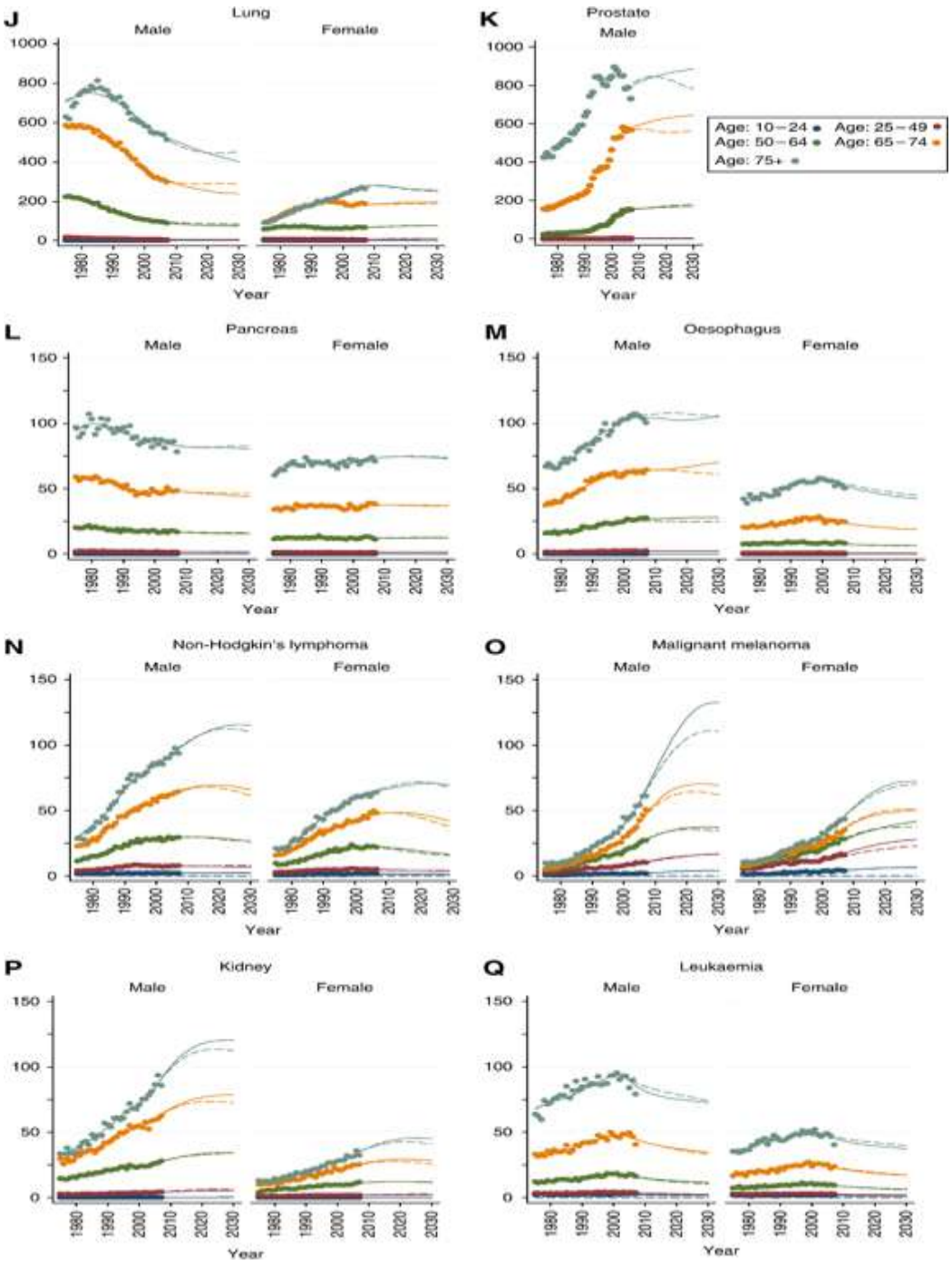


Figure B1: Observed and projected cancer incidence rates per 100,000 for various cancers. Rates are age standardised within each of the five age-bands. Projections using restricted cubic splines up to the year 2030 by Mistry et al., (2011).

Appendix B2: Lung Cancer Incidence in UK by Cancer Research UK (2009).

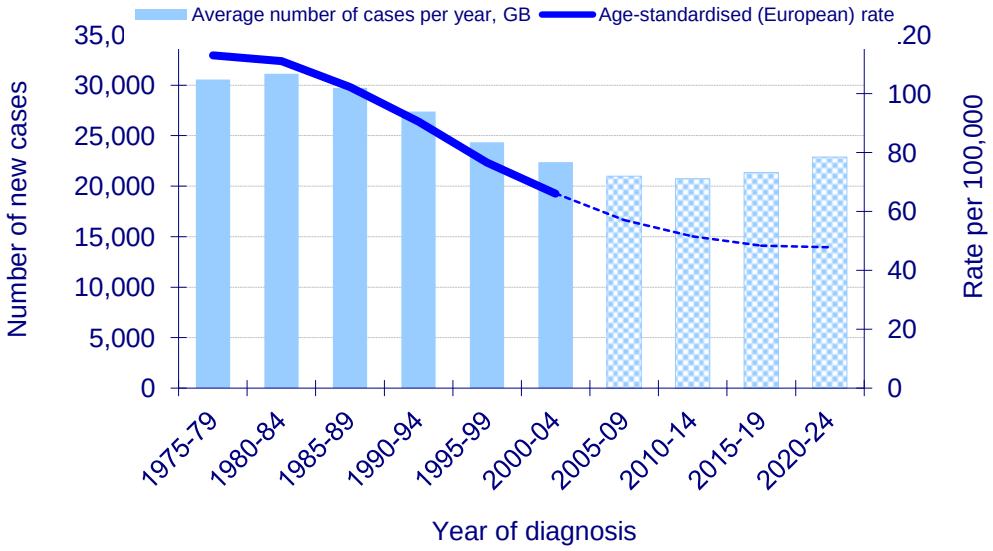


Figure B2: Males lung cancer incidence prediction to 2024, age-standardised rate and number of new cases, UK, from 1975 to 2004.

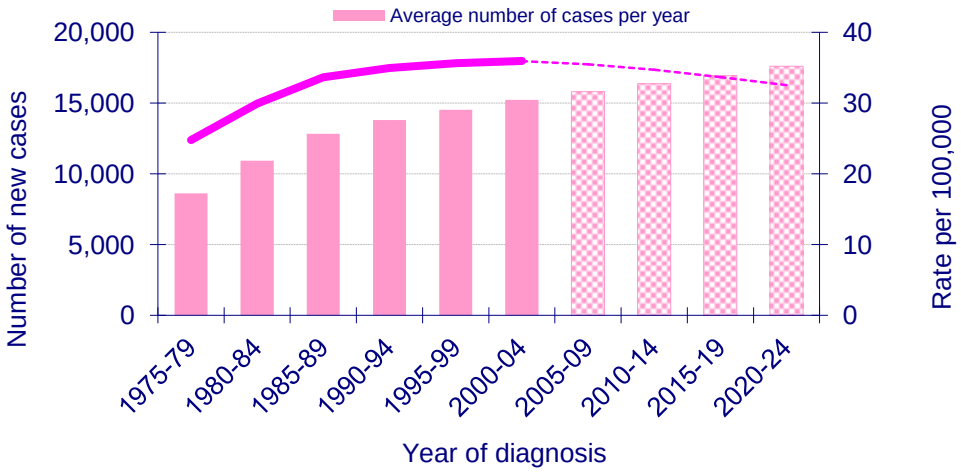


Figure B3: Females lung cancer incidence prediction to 2024, age-standardised rate and number of new cases, UK, from 1975 to 2004.

Appendix B3: Smoking Prevalence Among Males and Females in the UK According to Scottish Intercollegiate Guidelines Network, SIGNV(2005).

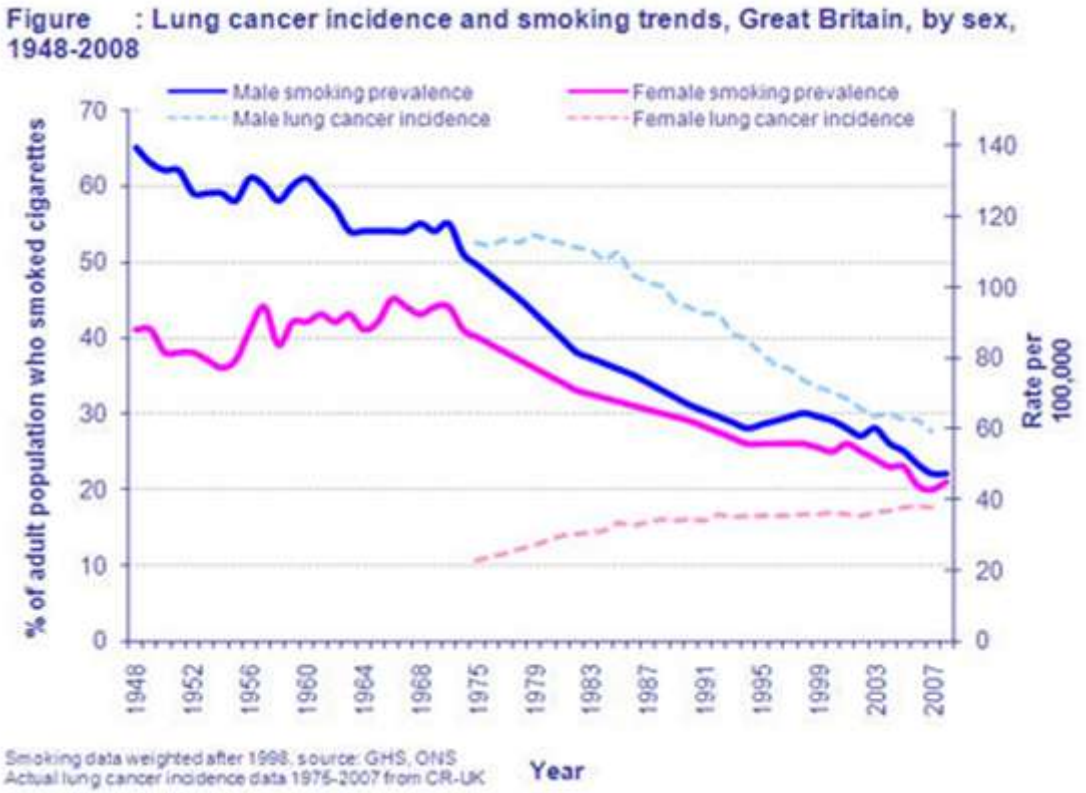


Figure B4: Trends in smoking prevalence in Britain from 1948 to 2008.

Appendix C: R Commands Used in Bayesian Dynamic APC Models

##AGE-PERIOD-COHORT MODEL##

```
model
{
### Fit of the Age, Period and Cohort
for (n in 1:N) {
deaths[n] ~ dpois(mu[n]);
# Modelling rate
log(mu[n]) <- log(pyr[n])+alpha[age[n]]+beta[period[n]]+gamma[cohort[n]];
}
#PERIOD EFFECTS:prior standard deviation
taup<-K.s*pow(sigmap,-2)
sigmap ~ dunif(0.01,1);
####PERIOD REF 1994
betamean[1]<-0.0;
betaprec[1]<-taup;
betamean[2]<-0.0;
betaprec[2]<-taup;
for (j in 3:J){
betamean[j]<-2*beta[j-1]-beta[j-2];
betaprec[j]<-taup;
}
#Corner constraint on the first period
beta[1]<-0
for (j in 2:J){
beta[j] ~ dnorm(betamean[j],betaprec[j]);
}
#COHORT EFFECTS:prior standard deviation
tauc<-K.s*pow(sigmac,-2)
sigmac ~ dunif(0.01,1);
####COHORT REF 1994
gammamean[1]<-0.0;
gammaprec[1]<-tauc;
gammamean[2]<-0.0;
gammaprec[2]<-tauc;
for (k in 3:K){
gammamean[k]<-2*gamma[k-1]-gamma[k-2];
gammaprec[k]<-tauc;
}
#Corner constraint on the first cohort
gamma[1]<-0
for (k in 2:K){
gamma[k] ~ dnorm(gammamean[k],gammaprec[k]);
}
#### AGE CONSTRAINED ON THE 2nd ORDER DIFFERENCES
alphamean[1] <- 2*alpha[2] - alpha[3];
alphamean[2] <- (2*alpha[1] + 4*alpha[3] - alpha[4])/5;
for (i in 3:(I-2)){
alphamean[i] <- (4*alpha[i-1] + 4*alpha[i+1]- alpha[i-2]
- alpha[i+2])/6;
}
alphamean[I-1] <- (2*alpha[I] + 4*alpha[I-2] - alpha[I-3])/5;
alphamean[I] <- 2*alpha[I-1] - alpha[I-2];
for (i in 1:I){
alphaprec[i] <- taua;
}
for (i in 1:I){
alpha[i] ~ dnorm(alphamean[i],alphaprec[i]);
}
#AGE EFFECTS:prior standard deviation
taua<-K.s*pow(sigmaa,-2);
sigmaa ~ dunif(0.01,1);
}
```

##AGE-PERIOD PREDICTION MODEL##

```
model
{

### Fit of the Age and Period
for (n in 1:N-M*I) {
deaths[n] ~ dpois(mu[n]);
# Modelling rate
log(mu[n]) <- log(pyr[n])+alpha[age[n]]+beta[period[n]];
}
#Modelling projections
for (i in 1:M*I) {
log(pred.mu[i])<-log(pyr[indx[i]])+alpha[age[indx[i]]]+beta[period[indx[i]]];

pred.rate[i]<-100000*pred.mu[i]/pyr[indx[i]];
}
#PERIOD EFFECTS:prior standard deviation
taup<-K.s*pow(sigmap,-2)
sigmap ~ dunif(0.01,1);


####PERIOD REF 1994
betamean[1]<-0.0;
betaprec[1]<-taup;
betamean[2]<-0.0;
betaprec[2]<-taup;
for (j in 3:J){
betamean[j]<-2*beta[j-1]-beta[j-2];
betaprec[j]<-taup;
}
#Corner constraint on the first period
beta[1]<-0
for (j in 2:J){
beta[j] ~ dnorm(betamean[j],betaprec[j]);
}


#### AGE CONSTRAINED ON THE 2nd ORDER DIFFERENCES
alphamean[1] <- 2*alpha[2] - alpha[3];
alphamean[2] <- (2*alpha[1] + 4*alpha[3] - alpha[4])/5;
for (i in 3:(I-2)){
alphamean[i] <- (4*alpha[i-1] + 4*alpha[i+1]- alpha[i-2]
- alpha[i+2])/6;
}
alphamean[I-1] <- (2*alpha[I] + 4*alpha[I-2] - alpha[I-3])/5;
alphamean[I] <- 2*alpha[I-1] - alpha[I-2];
for (i in 1:I){
alphaprec[i] <- taua;
}
for (i in 1:I){
alpha[i] ~ dnorm(alphamean[i],alphaprec[i]);
}
#AGE EFFECTS:prior standard deviation
taua<-K.s*pow(sigmaa,-2);
sigmaa ~ dunif(0.01,1);
}
```

Appendix D: Cases of Lung Cancer Mortality in KSA from 1994-2009 Prepared in the Lexis Diagram.

Age group		Period																
		1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	
25-29	Cohort	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	
	No. Deaths	0	1	0	1	2	0	2	1	1	1	0	0	3	1	3	1	
	No. at Risk	1822847	1881980	1941114	1964394	1987675	1859423	1731173	1857392	1780986	2237566	2293601	2370440	2317226	2305147	2373137	2568596	
30-34	Cohort	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	
	No. Deaths	1	0	2	3	0	4	2	2	2	1	3	3	6	1	3	3	
	No. at Risk	1641189	1674611	1708035	1765321	1822609	1809048.5	1795488	1878914	1868232	2073591	2125101	2200150	2358112	2373924	2441873	2711324	
35-39	Cohort	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	
	No. Deaths	1	1	1	2	4	4	3	3	5	9	1	7	0	5	1	3	
	No. at Risk	1380179	1380218	1380259	1410919	1441582	1487554	1533529	1579853	1672457	1810594	1855550	1921262	2103898	2095754	2155341	2415945	
40-44	Cohort	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	
	No. Deaths	0	2	0	1	3	4	5	1	5	10	8	11	9	9	12	9	
	No. at Risk	809201	846468.5	883736	903001	922269	1017578	1112889	1149311	1298844	1384324	1418761	1468411	1570478	1566299	1611296	1791422	
45-49	Cohort	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	
	No. Deaths	1	3	2	2	7	4	6	9	7	16	6	14	20	20	14	15	
	No. at Risk	652534	636943	621354	628734	636136	684401	732647	777244	914000	977374	1001774	1036036	1142031	1137094	1170141	1287787	
50-54	Cohort	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	
	No. Deaths	0	0	3	15	6	17	9	19	8	14	19	17	14	23	29	31	
	No. at Risk	413880	427539	441200	468697	496195	512272	528351	529298	607381	642121	658230	790018	782464	812666	836459	898635	
55-59	Cohort	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	
	No. Deaths	2	2	4	17	16	7	9	23	15	15	13	18	19	22	25	25	
	No. at Risk	302329	303591	304856	311453	318053	342604	367159	380974	424357	395775	405797	418375	490973	527699	543289	579341	
60-64	Cohort	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	
	No. Deaths	2	2	4	8	20	12	16	22	32	28	32	33	35	41	40	30	
	No. at Risk	271995	280954	289916	298388	306862	290010	273161	273001	301237	307655	315512	324678	324710	355397	366063	381876	
65-69	Cohort	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	
	No. Deaths	2	1	6	7	13	14	20	14	16	25	19	37	42	41	22	47	
	No. at Risk	141917	147545	153177	157500	161825	186445	211066	216913	218236	223245	228982	235308	210190	237744	244791	252319	
70-74	Cohort	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
	No. Deaths	0	1	4	3	14	12	15	22	20	20	24	33	45	39	34	27	
	No. at Risk	140044	143853	147666	151924	156183	170221	184260	185989	180654	173917	178393	183258	173976	185909	191524	197137	
75+	Cohort	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
	No. Deaths	1	0	1	16	13	28	22	23	24	23	34	47	43	53	59	52	
	No. at Risk	179804	191264	202724	208490	214239	228374	242491	243133	245071	219639	225302	231354	228046	232262	239741	265974	

Appendix F: Data for the Research Project.

Table F1: Cases of lung cancer among Saudi males from 1994 to 2009 for 16 age groups.

	Year of diagnosis (1994-2009)															
Age	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09
0 - 4	0	0	0	0	3	0	0	1	0	0	2	0	0	1	0	1
5-9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
10-14	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
15-19	0	1	2	0	0	0	0	0	0	0	2	0	0	1	0	0
20-24	1	1	0	0	1	0	1	1	0	0	1	0	1	2	0	1
25-29	1	1	0	2	2	1	0	1	0	1	0	1	1	2	2	2
30-34	1	3	5	2	4	4	2	2	2	3	4	3	2	1	3	4
35- 39	6	7	6	8	3	5	3	5	5	5	4	10	1	10	1	2
40-44	8	9	2	8	4	8	5	2	5	8	11	14	6	15	11	8
45-49	13	11	9	9	15	4	8	10	7	4	15	17	16	10	17	14
50- 54	20	16	17	18	19	19	8	18	13	18	26	21	15	26	35	27
55- 59	29	27	18	25	27	13	19	24	18	13	23	18	24	25	29	26
60- 64	36	37	36	26	35	27	31	28	44	28	45	42	30	49	41	36
65- 69	25	34	32	21	34	37	28	16	31	36	24	45	47	58	44	57
70- 74	29	31	28	21	45	28	30	28	27	28	31	46	43	47	36	41
75+	39	30	41	25	27	40	33	11	22	40	45	44	46	75	66	56
All	208	208	196	165	219	186	168	169	175	183	233	261	232	323	285	276

Table F2: Population (thousands) of Saudi males at risk from 1994 to 2009 for 16 age groups.

	Time period (1994-2009)															
Age	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09
0 - 4	1125	1175	1225	1286	1347	1306	1265	1338	1147	1039	1066	1093	1151	1157	1193	1209
5-9	1110	1160	1210	1257	1304	1248	1192	1292	1124	1099	1127	1156	1054	1085	1119	1134
10-14	1023	1022	1021	1072	1123	1092	1062	1072	1056	1055	1082	1109	993	1037	1069	1084
15-19	777	779	780	796	811	857	902	902	940	925	949	973	947	972	1002	1016
20-24	580	589	597	621	646	664	683	683	748	741	760	779	921	905	934	946
25-29	526	526	527	518	510	478	446	546	556	707	725	744	791	764	788	799
30-34	370	370	370	386	402	417	433	465	454	555	569	584	641	648	668	678
35- 39	312	320	328	326	325	352	380	390	379	480	493	505	538	529	546	553
40-44	123	169	216	219	221	266	311	321	331	401	412	422	441	439	453	459
45-49	187	187	186	188	189	212	235	245	273	305	313	321	358	352	363	368
50- 54	110	109	107	142	177	181	184	185	203	217	222	338	266	274	283	278
55- 59	107	107	107	126	144	144	143	143	155	142	146	150	181	202	208	211
60- 64	11	59	107	132	157	131	105	106	113	123	127	130	141	150	155	158
65- 69	55	56	56	72	87	97	106	116	103	101	103	106	76	103	106	108
70- 74	62	61	61	70	80	87	94	94	94	75	77	79	83	89	92	93
75+	83	84	84	98	113	123	133	134	132	113	116	119	83	116	120	122

Table F3: Cases of lung cancer among non-Saudi males from 1994 to 2009 for 16 age groups.

	Year of diagnosis (1994-2009)															
Age	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09
0 - 4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5-9	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
10-14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15-19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20-24	1	1	0	0	1	0	0	0	0	0	0	1	0	1	0	1
25-29	0	0	0	1	0	0	4	0	2	2	1	0	0	0	0	0
30-34	0	0	3	1	2	0	1	0	2	2	1	0	3	0	3	3
35- 39	4	4	2	1	7	3	3	4	3	3	2	5	2	4	5	5
40-44	5	5	8	6	11	7	6	3	6	6	2	10	8	5	6	6
45-49	14	14	10	5	9	4	4	7	7	7	6	13	10	14	6	6
50- 54	15	15	15	12	9	6	9	7	11	12	9	8	15	15	18	16
55- 59	21	20	13	11	7	2	4	6	11	11	8	14	11	19	12	10
60- 64	19	17	16	7	11	7	4	12	4	4	9	6	16	17	15	10
65- 69	9	9	12	4	4	1	6	5	13	14	7	8	12	9	10	9
70- 74	5	5	7	3	1	3	5	1	3	3	2	5	7	5	8	7
75+	7	7	5	5	1	3	2	0	1	0	4	7	5	7	2	7
All	100	97	92	56	63	37	48	45	63	65	51	77	90	96	85	80

Table F4: Population (thousands) of non-Saudi males at risk from 1994 to 2009 for 16 age groups.

	Time period (1994-2009)															
Age	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09
0 - 4	232	237	242	242	243	241	240	251	248	218	223	234	234	237	243	306
5-9	213	214	215	217	218	229	241	252	278	202	207	217	224	226	232	292
10-14	191	161	131	143	155	174	193	202	240	173	177	185	183	185	190	239
15-19	95	95	95	95	96	108	120	126	154	147	150	157	140	142	145	183
20-24	291	296	301	305	310	257	205	214	167	291	298	312	206	208	213	269
25-29	629	696	763	748	732	619	506	530	455	612	627	656	546	553	567	714
30-34	650	675	699	728	757	704	651	681	713	693	710	743	770	779	799	1006
35- 39	598	589	580	589	597	572	547	573	679	659	675	706	780	789	809	1019
40-44	380	374	368	376	383	376	369	386	486	500	512	536	560	567	581	732
45-49	230	220	210	208	205	209	213	223	293	330	338	354	373	377	386	487
50- 54	160	150	140	125	110	112	115	120	158	189	193	202	222	225	230	290
55- 59	61	56	51	49	48	54	60	63	82	83	85	89	107	109	111	140
60- 64	31	27	24	24	25	27	29	30	40	39	40	41	42	43	44	55
65- 69	9	9	9	9	9	11	12	13	18	16	17	17	17	18	18	23
70- 74	6	6	6	6	6	7	7	7	8	10	10	10	11	11	11	14
75+	6	6	6	6	6	7	8	9	9	9	9	9	12	13	13	16

Table F5: Cases of lung cancer among Saudi females from 1994 to 2009 for 16 age groups.

	Year of diagnosis (1994-2009)															
Age	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09
0 - 4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5-9	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1
10-14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15-19	0	1	0	1	0	0	0	0	0	1	1	1	1	0	0	1
20-24	0	1	0	0	0	0	0	1	0	0	3	0	0	0	2	0
25-29	1	0	0	2	1	0	0	1	1	2	1	1	4	3	1	3
30-34	1	1	5	3	0	0	0	0	2	3	0	2	1	0	2	2
35- 39	4	3	1	2	3	2	0	1	2	2	1	3	1	2	1	2
40-44	4	2	1	4	2	3	2	2	4	4	1	10	6	4	3	7
45-49	5	6	4	2	2	3	2	5	3	2	5	11	8	10	8	8
50- 54	5	4	7	8	5	6	9	2	3	8	5	8	5	10	4	20
55- 59	4	9	6	5	7	8	4	5	6	6	10	7	6	6	10	7
60- 64	6	8	5	9	6	10	5	9	5	9	13	13	13	7	15	9
65- 69	12	9	4	6	6	3	9	7	6	6	7	3	11	19	11	19
70- 74	3	5	10	5	3	3	6	6	3	5	6	11	11	8	11	6
75+	8	13	7	11	11	14	11	12	8	12	10	16	13	20	14	18
All	54	62	51	58	46	52	48	51	44	59	63	86	80	89	82	103

Table F6: Population (thousands) of Saudi females at risk from 1994 to 2009 for 16 age groups.

	Time period (1994-2009)															
Age	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09
0 - 4	1125	1197	1268	1285	1302	1243	1185	1285	1096	1028	1055	1081	1128	1123	1159	1516
5-9	1110	1169	1227	1256	1285	1218	1150	1236	1076	1084	1113	1141	1067	1073	1106	1099
10-14	1121	1089	1057	1071	1085	1058	1031	1041	1018	1126	1156	1185	1007	1037	1070	1063
15-19	770	767	764	796	829	810	792	892	917	915	939	963	954	952	982	976
20-24	587	598	608	621	635	643	652	752	787	767	787	806	931	880	908	902
25-29	517	498	480	520	559	590	621	616	630	684	701	719	756	762	786	781
30-34	380	379	378	386	393	437	481	491	493	561	575	590	650	646	666	663
35- 39	314	310	306	327	348	375	402	403	393	486	498	511	542	532	548	545
40-44	216	212	208	219	229	277	324	328	340	364	373	382	436	426	440	437
45-49	186	182	178	188	198	214	230	252	280	270	278	285	339	334	345	343
50- 54	118	142	166	174	182	191	199	193	212	194	199	204	249	267	275	274
55- 59	124	130	136	125	114	130	146	156	167	149	153	157	177	191	197	196
60- 64	118	133	148	131	114	121	128	125	136	131	134	138	135	147	152	151
65- 69	71	77	82	71	60	73	87	82	91	99	101	104	75	108	111	111
70- 74	67	71	75	70	65	70	75	76	71	83	85	87	72	78	80	80
75+	84	95	106	98	89	93	96	95	97	91	93	96	89	116	119	119

Table F7: Cases of lung cancer among non-Saudi females from 1994 to 2009 for 16 age groups.

	Year of diagnosis (1994-2009)															
Age	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09
0 - 4	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
5-9	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
10-14	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
15-19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20-24	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
25-29	1	0	0	0	2	0	0	0	1	1	1	0	0	0	1	0
30-34	5	3	2	0	2	0	0	0	3	3	1	1	5	0	1	2
35-39	0	2	3	1	1	2	1	0	0	0	1	3	1	2	0	0
40-44	3	1	1	2	3	0	0	0	3	3	0	3	6	1	1	4
45-49	1	2	2	2	0	4	1	2	2	2	1	0	2	5	1	3
50- 54	3	2	2	3	0	3	0	1	3	1	3	3	1	5	3	4
55- 59	3	5	3	3	1	2	3	0	1	3	1	5	3	4	1	3
60- 64	2	3	5	4	0	3	1	2	0	0	1	1	5	2	1	3
65- 69	3	2	2	2	1	2	0	2	1	1	1	6	3	1	3	6
70- 74	1	0	0	0	1	1	0	2	0	1	1	1	3	4	1	3
75+	1	0	0	0	0	0	0	1	1	0	0	5	0	2	4	3
All	24	20	20	18	11	19	6	12	15	15	12	31	29	26	17	31

Table F8: Population (thousands) of non-Saudi females at risk from 1994 to 2009 for 16 age groups.

	Time period (1994-2009)															
Age	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09
0 - 4	219	220	221	227	233	231	229	240	219	211	216	226	230	233	239	282
5-9	234	239	244	229	214	232	249	261	247	193	198	207	209	212	216	257
10-14	150	167	183	167	150	172	195	204	216	164	168	176	176	178	183	216
15-19	94	95	96	96	96	108	120	126	142	139	142	149	137	139	143	169
20-24	115	118	120	121	122	118	115	121	111	156	160	167	148	150	154	182
25-29	151	161	171	178	186	172	158	165	140	235	240	251	224	227	233	275
30-34	240	250	260	266	271	251	231	242	208	265	271	284	297	301	309	365
35- 39	156	161	166	169	171	188	205	214	222	186	190	199	243	246	252	299
40-44	90	91	91	90	88	99	109	114	142	119	122	128	133	134	138	163
45-49	49	48	47	45	44	49	54	56	67	71	73	76	73	74	76	90
50- 54	25	26	27	27	27	29	30	31	35	43	44	46	46	47	48	57
55- 59	10	11	11	12	12	15	18	19	20	21	22	23	26	26	27	32
60- 64	12	12	11	11	11	11	11	11	12	15	15	16	14	15	15	18
65- 69	6	6	5	5	5	6	6	6	7	8	8	8	9	9	9	11
70- 74	5	5	5	5	5	7	8	9	7	6	6	7	8	9	9	10
75+	6	6	6	6	6	6	6	6	7	7	7	8	8	8	8	10

Table F9: Male lung cancer cases per month from 1994 to 2009.

	Cases per month											
Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1994	47	21	20	28	24	32	24	20	19	19	19	13
1995	28	23	17	25	23	27	22	19	23	24	17	12
1996	35	21	19	24	19	24	18	17	25	21	13	14
1997	31	16	27	16	24	16	25	15	15	15	11	18
1998	23	16	31	22	32	21	31	30	25	22	14	16
1999	22	23	27	23	18	19	18	19	21	17	16	22
2000	26	24	20	15	25	24	23	27	21	17	18	12
2001	24	23	14	27	25	20	22	14	22	19	19	17
2002	34	20	31	24	22	30	20	14	16	24	15	27
2003	30	29	23	24	24	23	26	22	23	23	17	23
2004	27	29	27	27	39	28	28	17	28	16	30	30
2005	25	25	38	38	30	27	28	20	35	32	25	30
2006	26	29	40	36	35	32	29	21	23	20	15	28
2007	35	38	34	36	50	38	39	33	31	22	30	40
2008	36	38	33	31	34	34	32	30	27	30	33	24
2009	28	39	36	30	32	34	23	30	17	39	30	39

Table F10: Female lung cancer cases per month from 1994 to 2009.

	Cases per month											
Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1994	10	6	5	6	5	4	5	10	7	5	5	6
1995	9	7	5	6	3	9	5	6	4	10	4	12
1996	7	4	8	4	3	7	5	7	5	8	2	4
1997	12	7	5	5	6	8	7	2	2	6	3	7
1998	1	4	5	3	6	6	8	7	5	8	6	4
1999	7	6	8	9	11	4	4	5	4	2	3	9
2000	2	5	3	7	7	1	11	8	4	3	6	4
2001	6	5	7	6	7	3	8	4	6	5	8	5
2002	9	1	6	9	4	6	6	9	6	5	7	7
2003	10	5	4	9	8	6	5	9	7	8	9	6
2004	10	6	13	9	10	7	5	2	4	4	7	7
2005	10	8	13	10	11	11	6	11	13	9	7	9
2006	12	13	8	11	5	6	10	8	6	8	8	16
2007	9	9	14	5	9	14	7	9	10	11	10	15
2008	8	8	10	6	13	10	8	11	4	8	9	8
2009	8	18	14	8	13	14	17	8	12	7	12	10

Table F11: Male smoking population in 10,000 per month from 1994 to 2009.

	Male smoking population in ten thousand per month											
Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1994	196	196	195	195	194	194	194	193	193	193	192	192
1995	192	192	192	193	193	193	194	194	194	195	195	195
1996	196	196	197	197	198	199	199	200	200	201	202	202
1997	203	204	205	206	206	207	208	209	210	211	212	213
1998	214	213	212	211	209	208	207	206	205	204	203	202
1999	201	200	200	199	199	198	198	197	196	196	195	195
2000	194	195	196	196	197	197	198	198	199	200	200	201
2001	201	202	203	205	206	207	208	209	210	211	213	214
2002	215	216	217	218	219	220	221	222	223	224	225	226
2003	227	228	229	231	232	233	234	236	237	238	239	241
2004	242	244	246	248	250	252	254	256	257	259	261	263
2005	265	265	266	266	266	266	267	267	267	267	267	268
2006	268	270	273	275	278	280	283	285	288	290	293	295
2007	298	296	295	293	292	290	289	287	286	284	283	281
2008	279	282	285	288	290	293	296	298	301	304	307	309
2009	312	312	312	312	312	312	312	312	312	312	312	312

Table F12: Female smoking population in 10,000 per month from 1994 to 2009.

	Female smoking population in ten thousand per month											
Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1994	31.1	31.5	31.9	32.3	32.6	33	33.4	33.8	34.2	34.5	34.9	35.3
1995	35.7	35.5	35.4	35.2	35.1	34.9	34.8	34.6	34.5	34.3	34.2	34
1996	33.9	34	34.1	34.2	34.3	34.4	34.5	34.6	34.7	34.8	34.9	35
1997	35.1	35.1	35.1	35.1	35.1	35.1	35.1	35.1	35.1	35.1	35.1	35.1
1998	35.1	35.2	35.2	35.2	35.2	35.3	35.3	35.3	35.4	35.4	35.4	35.5
1999	35.5	35	34.5	34	33.6	33.1	32.6	32.1	31.6	31.1	30.7	30.2
2000	29.7	30.5	31.3	32.1	32.9	33.7	34.5	35.4	36.2	37	37.8	38.6
2001	39.4	39.3	39.1	39	38.9	38.7	38.6	38.5	38.3	38.2	38.1	38
2002	37.8	38.2	38.5	38.9	39.3	39.6	40	40.3	40.7	41.1	41.4	41.8
2003	42.1	42.3	42.4	42.5	42.6	42.8	42.9	43	43.1	43.3	43.4	43.5
2004	43.6	44.5	45.3	46.2	47	47.9	48.8	49.6	50.5	51.3	52.2	53
2005	53.9	53.9	53.9	53.9	54	54	54	54.1	54.1	54.1	54.1	54.2
2006	54.2	54.2	54.3	54.4	54.4	54.5	54.5	54.6	54.7	54.7	54.8	54.8
2007	54.9	55	55.1	55.2	55.3	55.4	55.5	55.6	55.7	55.8	55.9	56
2008	56.1	56.9	57.7	58.5	59.2	60	60.8	61.6	62.4	63.2	64	64.8
2009	65.5	65.5	65.5	65.5	65.5	65.5	65.5	65.5	65.5	65.5	65.5	65.5

Table F13: Total cases of lung cancer in KSA from 1994 to 2009 for 16 age groups.

	Year of diagnosis (1994-2009)															
Age	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09
0 - 4	1	0	0	0	3	1	0	1	0	0	2	0	0	1	0	1
5-9	0	0	2	0	0	1	0	0	1	0	0	0	1	0	0	2
10-14	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	0
15-19	0	2	2	1	0	0	0	0	0	1	3	1	1	1	0	1
20-24	2	3	0	1	2	0	1	2	0	0	5	1	1	3	2	2
25-29	3	1	0	5	5	1	4	2	4	6	3	2	5	5	4	5
30-34	7	7	15	6	8	4	3	2	9	11	6	6	11	1	9	11
35-39	14	16	12	12	14	12	7	10	10	10	8	21	5	18	7	9
40-44	20	17	12	20	20	18	13	7	18	21	14	37	26	25	21	25
45-49	33	33	25	18	26	15	15	24	19	15	27	41	36	39	32	31
50- 54	43	37	41	41	33	34	26	28	30	39	43	40	36	56	60	67
55- 59	57	61	40	44	42	25	30	35	36	33	42	44	44	54	52	46
60- 64	63	65	62	46	52	47	41	51	53	41	68	62	64	75	72	58
65- 69	49	54	50	33	45	43	43	30	51	57	39	62	73	87	68	91
70- 74	38	41	45	29	50	35	41	37	33	37	40	63	64	64	56	57
75+	55	50	53	41	39	57	46	24	32	52	59	72	64	104	86	84
All	385	387	359	297	339	293	270	255	296	323	359	452	431	534	469	490

Table F14: Person-years-at- risk (in thousands) in five-year age groups and one-year time period from 1994-2009.

	Time Period															
Age	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09
0 - 4	2702	2829	2957	3040	3124	3021	2918	3113	2710	2496	2560	2634	2743	2750	2834	3313
5-9	2666	2781	2896	2959	3022	2927	2832	3041	2725	2579	2645	2720	2553	2595	2674	2782
10-14	2485	2439	2392	2452	2512	2496	2480	2519	2530	2517	2582	2655	2358	2437	2512	2602
15-19	1736	1735	1734	1783	1832	1883	1934	2046	2152	2126	2180	2242	2178	2205	2273	2344
20-24	1574	1600	1626	1669	1712	1683	1655	1770	1813	1955	2005	2065	2206	2143	2209	2298
25-29	1823	1882	1941	1964	1988	1859	1731	1857	1781	2238	2294	2370	2317	2305	2373	2569
30-34	1641	1675	1708	1765	1823	1809	1795	1879	1868	2074	2125	2200	2358	2374	2442	2711
35- 39	1380	1380	1380	1411	1442	1488	1534	1580	1672	1811	1856	1921	2104	2096	2155	2416
40-44	809	846	884	903	922	1018	1113	1149	1299	1384	1419	1468	1570	1566	1611	1791
45-49	653	637	621	629	636	684	733	777	914	977	1002	1036	1142	1137	1170	1288
50- 54	414	428	441	469	496	512	528	529	607	642	658	790	782	813	836	899
55- 59	302	304	305	311	318	343	367	381	424	396	406	418	491	528	543	579
60- 64	272	231	290	298	307	290	273	273	301	308	316	325	332	355	366	382
65- 69	142	148	153	158	162	186	211	217	218	223	229	235	177	238	245	252
70- 74	140	144	148	152	156	170	184	186	181	174	178	183	174	186	192	197
75+	180	191	203	208	214	228	242	243	245	220	225	231	193	252	260	266

Table F15: Cases of lung cancer mortality for males from 1994 to 2009 for 16 age groups.

	Year of diagnosis (1994-2009)															
Age	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09
0 - 4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
5-9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10-14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15-19	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0
20-24	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0
25-29	0	1	0	0	1	0	2	1	0	1	0	0	2	0	1	0
30-34	0	0	1	1	0	3	2	2	2	1	2	2	4	1	1	2
35- 39	1	1	1	2	2	4	1	3	2	9	1	6	0	5	1	3
40-44	0	1	0	0	2	4	4	1	3	7	8	5	4	8	10	4
45-49	1	3	2	2	6	3	5	6	6	13	5	9	15	11	9	11
50- 54	0	0	2	11	4	15	6	18	6	14	14	11	12	13	26	17
55- 59	1	1	2	12	15	5	9	19	11	12	13	13	17	20	17	18
60- 64	1	2	4	6	18	9	15	20	30	23	24	28	24	35	28	28
65- 69	1	0	2	7	11	10	16	10	13	19	16	31	33	32	16	36
70- 74	0	1	3	3	12	11	14	17	17	16	20	25	35	35	28	22
75+	1	0	1	15	11	20	19	20	20	15	30	32	31	41	49	39
All	6	10	18	59	82	84	93	117	111	131	135	162	178	203	186	180

Table F16: Cases of lung cancer mortality for females from 1994 to 2009 for 16 age groups.

	Year of diagnosis (1994-2009)															
Age	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09
0 - 4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5-9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10-14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15-19	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
20-24	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0
25-29	0	0	0	1	1	0	0	0	1	0	0	0	1	1	2	1
30-34	1	0	1	2	0	1	0	0	0	0	1	1	2	0	2	1
35- 39	0	0	0	0	2	0	2	0	3	0	0	1	0	0	0	0
40-44	0	1	0	1	1	0	1	0	2	3	0	6	5	1	2	5
45-49	0	0	0	0	1	1	1	3	1	3	1	5	5	9	5	4
50- 54	0	0	1	4	2	2	3	1	2	0	5	6	2	10	3	14
55- 59	1	1	2	5	1	2	0	4	4	3	0	5	2	2	8	7
60- 64	1	0	0	2	2	3	1	2	2	5	8	5	11	6	12	2
65- 69	1	1	4	0	2	4	4	4	3	6	3	6	9	9	6	11
70- 74	0	0	1	0	2	1	1	5	3	4	4	8	10	4	6	5
75+	0	0	0	1	2	8	3	3	4	8	4	15	12	12	10	13
All	4	3	9	16	16	22	16	22	25	32	26	59	61	54	58	63

Table F17: Male lung cancer mortality cases per month from 1994 to 2009.

	Male deaths per month											
Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1994	0	0	0	1	3	0	0	1	0	0	0	1
1995	1	1	1	2	0	0	0	2	0	1	1	1
1996	2	1	4	2	0	2	1	0	3	0	1	2
1997	5	3	8	3	8	3	6	4	5	4	3	6
1998	8	2	13	6	10	5	9	8	5	7	5	7
1999	5	11	7	7	8	5	5	5	11	6	8	6
2000	8	6	6	4	13	7	9	10	7	6	7	10
2001	10	14	8	9	13	11	10	5	9	9	10	6
2002	17	9	10	10	8	13	7	4	5	9	8	10
2003	11	12	11	12	8	12	14	10	12	11	7	14
2004	10	10	15	14	16	12	13	6	16	3	8	11
2005	11	10	20	21	14	8	16	10	15	14	10	11
2006	12	7	19	17	14	11	7	9	7	5	7	6
2007	16	19	9	15	20	21	15	18	9	17	11	12
2008	16	19	20	15	13	14	16	11	10	14	10	11
2009	16	15	15	11	15	17	10	9	10	18	16	17

Table F18: Female lung cancer mortality cases per month from 1994 to 2009.

	Female deaths per month											
Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1994	0	0	0	1	0	0	0	1	0	0	2	0
1995	1	1	0	0	0	1	0	0	0	0	0	0
1996	1	1	1	2	0	0	2	0	1	0	0	1
1997	1	3	1	1	2	3	1	0	0	1	1	0
1998	0	2	2	1	0	1	3	1	1	2	3	0
1999	4	1	1	4	3	3	1	1	1	1	2	0
2000	0	2	1	1	3	0	2	4	1	0	1	1
2001	1	1	3	1	2	2	5	2	1	1	2	2
2002	3	1	1	4	2	2	2	4	3	0	1	2
2003	3	2	2	2	2	1	3	4	3	5	3	1
2004	2	3	2	4	2	4	4	1	2	2	0	0
2005	4	5	9	4	2	4	4	7	8	4	4	3
2006	7	4	3	8	3	4	6	4	5	4	4	11
2007	4	2	5	3	6	7	5	5	6	4	4	6
2008	6	5	5	2	4	2	7	6	2	4	4	3
2009	3	3	6	8	7	4	4	4	7	3	6	4

Table F19: Male population (thousands) at risk forecast from 2010 to 2020 for 16 age groups.

	Time period										
Age	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
0 - 4	1507	1498	1490	1482	1473	1465	1450	1435	1420	1405	1391
5-9	1451	1477	1504	1531	1558	1585	1572	1559	1546	1533	1520
10-14	1339	1352	1365	1379	1392	1405	1422	1438	1455	1471	1488
15-19	1209	1218	1228	1237	1246	1256	1283	1310	1337	1364	1391
20-24	1238	1259	1281	1302	1324	1345	1367	1389	1411	1433	1455
25-29	1531	1548	1564	1581	1598	1614	1628	1641	1655	1668	1682
30-34	1694	1708	1722	1736	1750	1764	1806	1847	1889	1931	1973
35- 39	1639	1705	1772	1839	1906	1973	1973	1973	1973	1973	1973
40-44	1271	1345	1420	1495	1570	1644	1710	1776	1841	1907	1973
45-49	892	929	966	1003	1039	1076	1159	1241	1323	1405	1488
50- 54	614	653	691	730	769	807	846	885	924	963	1003
55- 59	397	443	489	536	582	628	651	674	697	721	744
60- 64	254	293	332	371	409	448	475	502	529	555	582
65- 69	136	139	142	144	147	149	191	232	273	314	356
70- 74	106	109	112	114	117	120	122	123	125	127	129
75+	380	334	288	242	196	149	145	141	137	133	129

Table F20: Female population (thousands) at risk forecast from 2010 to 2020 for 16 age groups.

	Time period										
Age	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
0 - 4	1433	1427	1422	1416	1411	1405	1389	1373	1358	1342	1326
5-9	1376	1394	1412	1429	1447	1465	1444	1422	1401	1380	1358
10-14	1264	1244	1225	1205	1186	1166	1172	1178	1184	1190	1197
15-19	1099	1053	1006	960	914	867	862	857	851	846	841
20-24	1075	1069	1063	1058	1052	1046	999	951	904	856	809
25-29	1090	1124	1157	1190	1223	1256	1231	1206	1181	1157	1132
30-34	1047	1071	1095	1118	1142	1166	1204	1243	1281	1320	1358
35- 39	876	910	944	978	1012	1046	1057	1068	1078	1089	1100
40-44	619	638	658	678	698	718	749	780	811	842	873
45-49	469	500	532	564	596	628	645	661	678	695	711
50- 54	371	410	450	489	529	568	597	625	654	683	711
55- 59	254	281	308	335	362	389	414	440	466	492	517
60- 64	190	212	234	255	277	299	310	322	333	344	356
65- 69	129	139	149	159	169	179	208	237	266	295	323
70- 74	95	100	105	110	115	120	122	123	125	127	129
75+	129	133	137	141	145	149	171	193	215	237	259

Appendix G: ARPDL Models with Few Number of Lags

G.0. Dynamic modelling with a few number of lags approach

Suppose that we want to estimate just a few number of lags in our model, say 6 lags for both the dependent (Y_{t-i}) and the independent ($X_{t-\mathrm{i}}$) variables. We use the same procedure as we did before for choosing the best lag length and order of the polynomial to select the best ARPDL model. Next, we check the validity of the chosen best-fit model by using the cross-validation procedure using the one step ahead out-of-sample forecasts. This will provide us with a yardstick to compare the presented models with high number of lags. This is presented through the following steps:

1- Choose the best lag length of the independent variable (k).

2- Choose the best order of the polynomial of the independent variable (r).

3- Choose the best lag length of the dependent variable (p).

4- Choose the best order of the polynomial of the dependent variable (q).

5- Run the ARPDL(p,q,k,r) model and check the model diagnostic plots.

6- Perform cross-validation using the one step ahead out-of-sample forecast from 2008 to 2009 and select the best fit model.

G.1. Choosing the Lag Length with OLS for the Independent Variable

We run the regression 6 times using different lags, starting from lag 6 to lag 1. Then, we checked where the fit of the models deteriorates significantly.

Table G1: Choosing the best lag length from OLS

	Lag					
coefficient of	6*	5	4	3	2	1
x_{t-1}	0.071	0.065	0.053	0.048	0.041	0.142
x_{t-2}	0.265	0.265	0.265	0.265	0.103	
x_{t-3}	-0.082	-0.082	-0.082	-0.172		
x_{t-4}	0.146	0.146	-0.097			
x_{t-5}	-0.174	-0.261				
x_{t-6}	-0.094					
sum of coefficient	0.132	0.134	0.139	0.141	0.144	0.142
$\bar{R}^2$	0.457	0.451	0.434	0.430	0.427	0.428
DW	1.84	1.81	1.80	1.76	1.75	1.80

The best lag length is 6 based on the highest adjusted R-squared. The main problem with the OLS estimates is that, no matter how many lags we include, the Durbin-Watson

(DW) test shows positive correlation (less than 2). From Table G1, DW suggests a typical symptom of collinearity and we should be estimating some more general dynamic models, allowing for autocorrelated errors. Thus, we use the polynomial distributed lag model.

G.2. Choosing the Degree of the Polynomial for the Independent Variable

Having determined the best lag length of the independent variable (x_{t-i}). The next step is to specify the order of the polynomial by starting with a high-degree polynomial and then we decrease it until we obtain a satisfactory fit. So we started with a polynomial of degree three and decreased it until we obtained a satisfactory fit as shown in Table G2.

Table G2: Choosing the degree of the polynomial.

	Equation								
	1			2*			3		
coefficient of	3rd order	t ratios	p-value	2nd order	t ratios	p-value	1st order	t ratios	p-value
Z_{0t}	0.00	-0.09	0.93	0.00	-0.08	0.93	0.02	9.21	0.00
Z_{1t}	-0.09	-1.50	0.14	-0.04	-3.91	0.00	-0.04	-3.94	0.00
Z_{2t}	0.01	0.54	0.59	0.01	0.53	0.60			
Z_{3t}	0.01	0.86	0.39						
$\bar{R}^2$	0.45			0.46			0.46		
σ^2	6432.09			6462.73			6474.40		
DW	1.870			1.873			1.856		

We compare the adjusted R-squared values for the three models and their corresponding DW statistics to select the best order for the polynomial. From Table G2, the 2nd-order polynomial is appropriate due to its adjusted R-squared and DW statistic (close to 2). Hence, the bet model of the polynomial distributed lag models is PDL(6,2).

G.3. Choosing the Lag Length of Y_t from OLS

The best lag length of Y_t is as shown (starred) in Table G3. We ran the regression 6 times using different lags of y_t, starting from lag 6 to lag 1. Then, we checked where the fit of the models deteriorates significantly.

Table G3: Choosing the best lag length of Y_t from ordinary least squares.

	Lag					
coefficient of	6	5	4	3*	2	1
y_{t-1}	0.29	0.30	0.30	0.29	0.36	0.46
y_{t-2}	0.29	0.28	0.29	0.29	0.30	
y_{t-3}	0.11	0.11	0.13	0.14		
y_{t-4}	-0.03	-0.01	0.00			
y_{t-5}	0.01	0.05				
y_{t-6}	0.11					
$\bar{R}^2$	0.316	0.313	0.315	0.320	0.298	0.212
AIC	6.83	6.82	6.81	6.79	6.81	6.92

From Table G3, the appropriate lag length of Y_t is 3. This is due to the highest adjusted R-squared and lowest value of AIC.

G.4. Choosing the Degree of the Polynomial Y_t

Here, we started with a third-degree polynomial and decreased it until we obtained a satisfactory fit.

Table G4: Choosing the degree of the polynomial.

	Equation								
	1			2			3*		
coefficient of	3rd order	t ratios	p-value	2nd order	t ratios	p-value	1st order	t ratios	p-value
PDL01	0.29	3.53	0.00	0.26	4.86	0.00	0.23	8.13	0.00
PDL02	-0.12	-1.06	0.29	-0.07	-1.17	0.24	-0.10	-2.75	0.01
PDL03	-0.08	-0.71	0.48	-0.03	-0.64	0.52			
PDL04	0.03	0.49	0.63						
$\bar{R}^2$	0.315			0.319			0.321		
σ^2	8165.77			8177.92			8199.08		
AIC	6.81			6.80			6.79		

Therefore, the best order of the polynomial is 1. This is due to the highest adjusted R-squared and lowest value of AIC as shown (starred) in Table G4 above. Hence, the best model is ARPDL(3,1,6,2);

$$Y_t = \alpha + \sum_{i=1}^{6} (\gamma_0 i^0 + \gamma_1 i^1 + \gamma_2 i^2) X_{t-i} + \sum_{i=1}^{3} (\gamma_0 i^0 + \gamma_1 i^1) Y_{t-i} + \varepsilon_t \qquad \text{G.1}$$

Table G5: Results of the autoregressive polynomial distributed lag ARPDL(3,1,6,2) model.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-4.93	3.72	-1.32	0.18
Z_{0t}	0.01	0.04	0.33	0.74
Z_{1t}	-0.03	0.01	-3.14	0.00
Z_{2t}	0.00	0.01	0.12	0.90
Z_{3t}	0.02	0.04	0.52	0.60
Z_{4t}	-0.06	0.03	-1.84	0.06

R-squared	0.481866	Mean dependent var	30.85093
Adjusted R-squared	0.465152	S.D. dependent var	8.729412
S.E. of regression	6.384109	Akaike info criterion	6.582056
Sum squared resid	6317.311	Schwarz criterion	6.696891
Log likelihood	-523.8555	Hannan-Quinn criter.	6.628683
F-statistic	28.83006	Durbin-Watson stat	2.074427
Prob(F-statistic)	0.000000		

Note that the created variables from Z_{0t} to Z_{2t} refer to the lag of X_{t-i} whereas the variables from Z_{3t} to Z_{4t} refer to the lag of Y_{t-i}.

After fitting the dynamic regression model it is important to determine whether all the necessary model assumptions are valid before performing any forecast. If there are any violations, subsequent inferential procedures may be invalid resulting in faulty conclusions. Therefore, it is crucial to perform appropriate model diagnostics.

The fitted model is shown in Figure G1 together with residual diagnostic plots. This is followed by the distribution of the series in the histogram with a complement of standard descriptive statistics displayed along with the histogram (see Figure G2). The p-value (p=0.31) of the Jarque-Bera test is not less than 0.05 for a 5% significance level and hence we do not reject the null hypothesis that the model is normally distributed. Figure G3 shows leverage plots of the residuals.

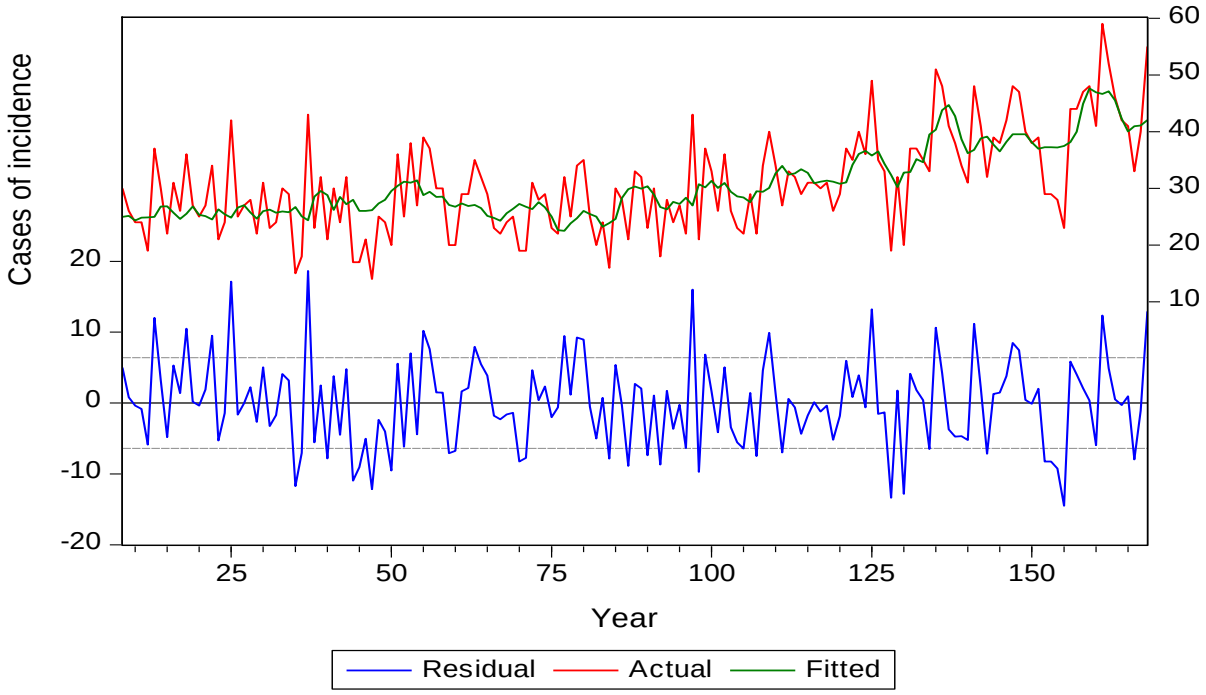


Figure G1: Fitted and residual plots for the best ARPDL(3,1,6,2) model of lung cancer cases per month from 1994 to 2009.

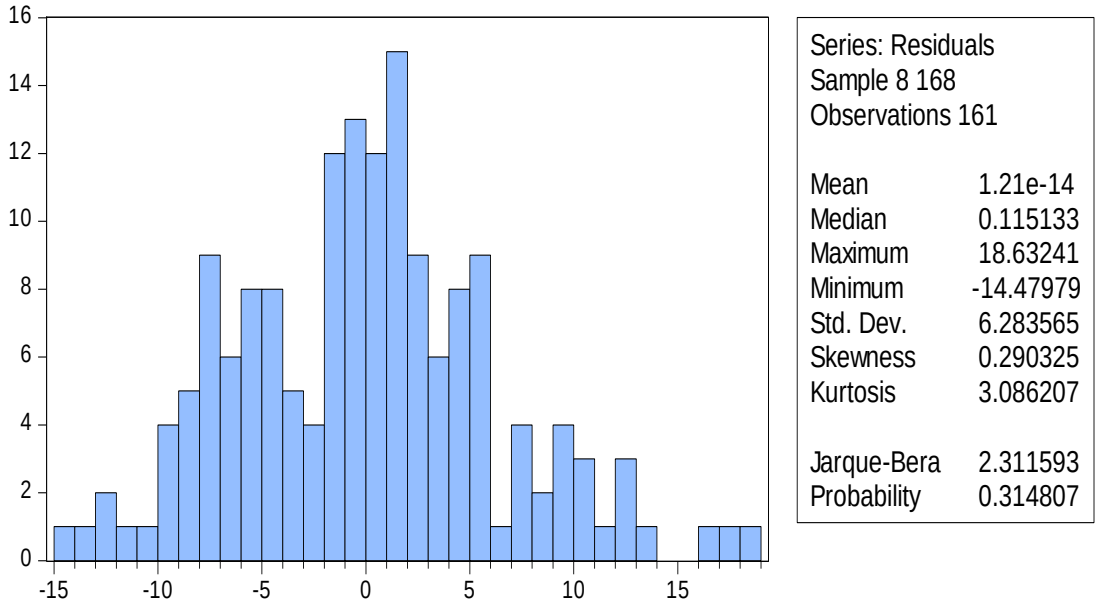


Figure G2: Residual diagnostic of the normality test of the best ARPDL(3,1,6,2) model of lung cancer cases per month from 1994 to 2009.

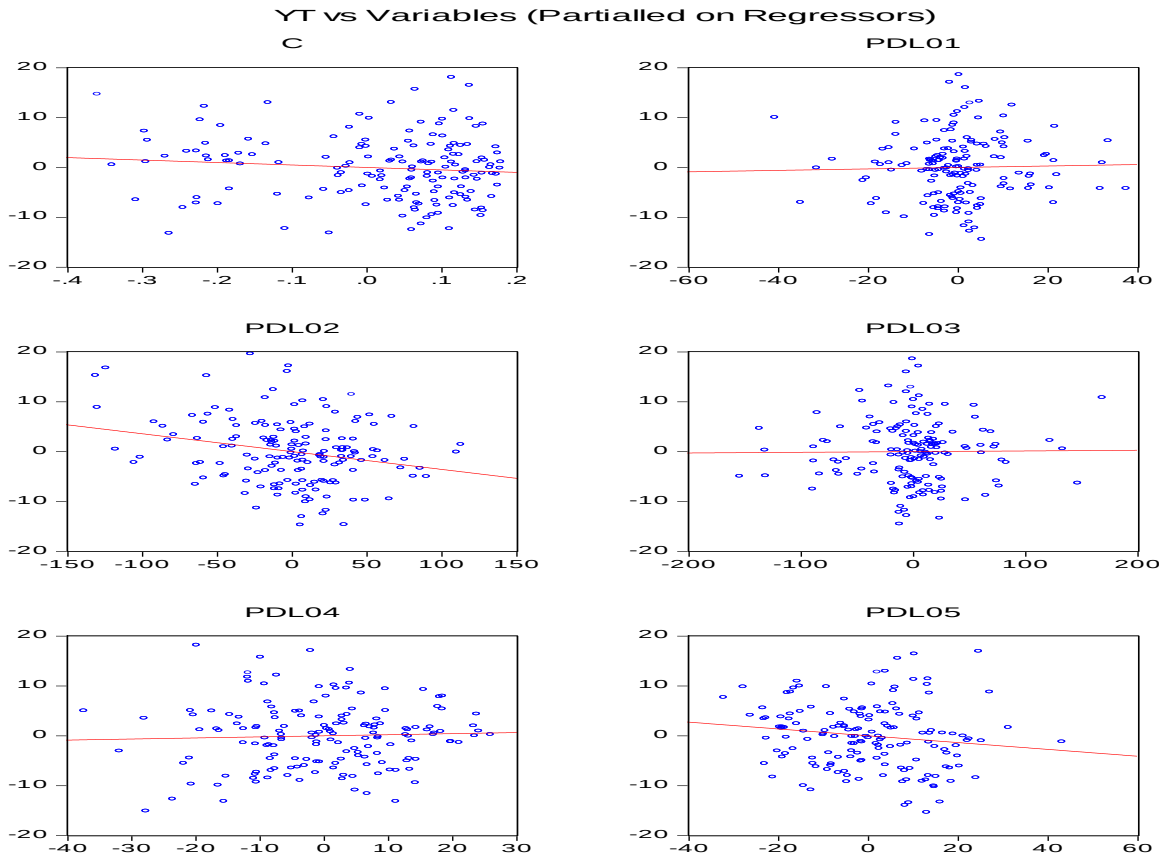


Figure G3: Leverage plots for the stability of diagnostics of the best OLS model of lung cancer cases per month from 1994 to 2009.

G.5. The Breusch-Godfrey Test for Serial Correlation

From Table G6, the values of both the LM-statistic and the F-statistic are quite low, indicating that we do not reject the null hypothesis and hence conclude that there is no significant serial correlation. Residuals generated from the model are not serially correlated because the p-values are not very small i.e. they are not less than 0.05 for a 5% significance level. Hence, we forecast this model and present the k-step ahead forecast as shown Figure G4.

Table G6: Results of Breusch-Godfrey LM test of ARPDL(3,1,6,2) model.

F-statistic	2.264395	Prob. F(5,155)	0.0507
Obs*R-squared	10.95970	Prob. Chi-Square(5)	0.0522

G.6. Results

The one step ahead out-of-sample forecast was performed on the data from 2008 to 2009 to check the validity of the ARPDL(3,1,6,2) model (Figure G4). Figure G5 shows the actual cases of lung cancer from 1994 to 2009 and the forecast value between 2008 and 2009.

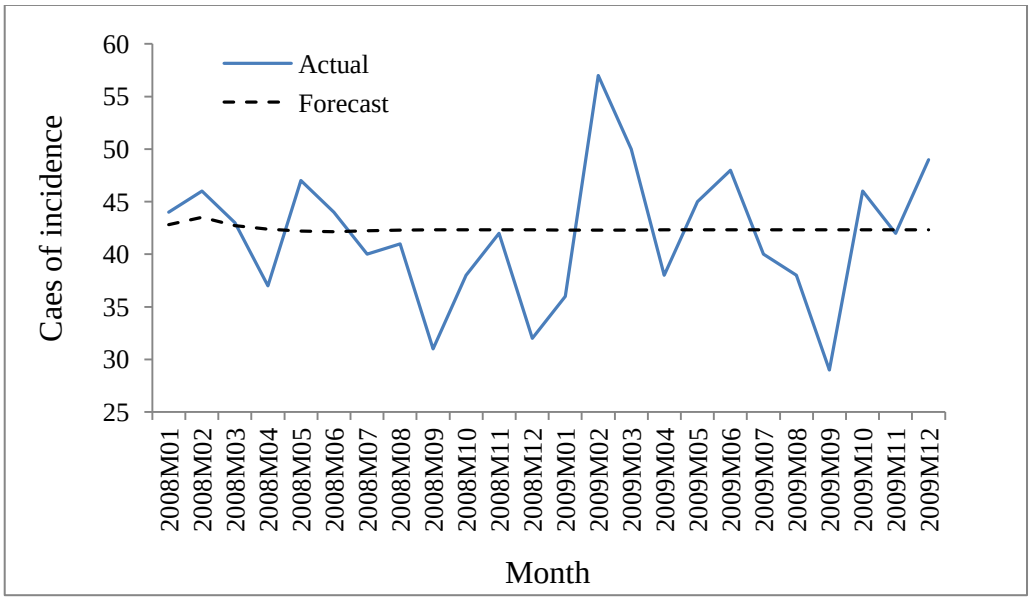


Figure G4: Actual and forecast ARPDL(3,1,6,2) model with 24 months ahead forecast of lung cancer cases per month from 2008 to 2009.

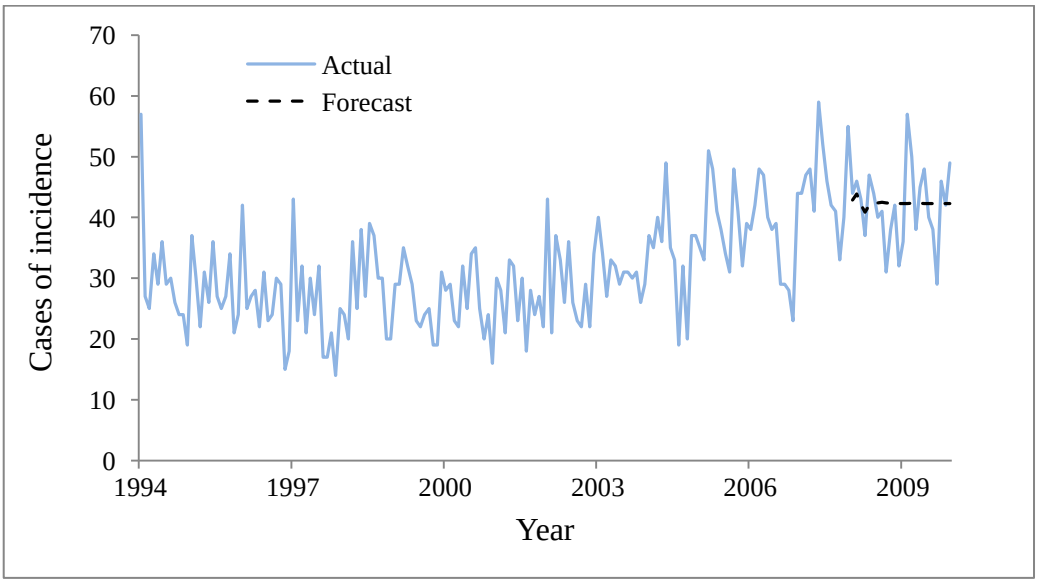


Figure G5: Actual and forecast ARPDL(3,1,6,2) model with 24 months ahead forecast of lung cancer cases per month from 1994 to 2009.

Appendix S: ARPDL Models with high Number of Lags

Here, we have decided to analyze the data between 2000 and 2007. Next, we perform cross-validation of the model by using the one step ahead out-of-sample forecasts for the next 24 month through the following steps:

S.1. Choosing the Lag Length with OLS for the Independent Variable

We run the regression 24 times using different lags, starting from lag 24 to lag 1. Then, we checked where the fit of the models deteriorates significantly.

Table S1: Choosing the best lag length from OLS

	lag			
Model Statistics	24	23*	22	21
$\bar{R}^2$	0.54	0.55	0.50	0.51
DW	1.94	2.09	2.14	2.03
AIC	6.66	6.65	6.73	6.71

The best lag length is 23 regarding to the highest adjusted R-squared. From Table S1, DW suggests a typical symptom of collinearity and we should be estimating some more general dynamic models, allowing for autocorrelated errors. Thus, we use the polynomial distributed lag model.

S.2. Choosing the Degree of the Polynomial for the Independent Variable

Having determined the best lag length of the independent variable (x_{t-i}). The next step is to specify the degree of the polynomial by starting with a high-degree polynomial and then we decrease it until we obtain a satisfactory fit. So we started with a polynomial of degree six and decreased it until we obtained a satisfactory fit as shown in Table S2.

Table S2: Choosing the degree of the polynomial.

	Equation			
Model Statistics	6th* order	5th order	4th order	3rd order
$\bar{R}^2$	0.56	0.53	0.54	0.40
DW	2.17	2.15	2.15	1.64
AIC	6.45	6.49	6.47	6.72

From Table S2, the 6th-order polynomial is appropriate due to its highest adjusted R-squared and lowest value of AIC. Hence, the bet model of the polynomial distributed lag models is PDL(23,6).

S.3. Choosing the Lag Length of Y_t from OLS

The best lag length of Y_t is as shown (starred) in Table S3. We ran the regression 12 times using different lags of y_t, starting from lag 12 to lag 1. Then, we checked where the fit of the models deteriorates significantly.

Table S3: Choosing the best lag length of Y_t from ordinary least squares.

	lag				
Model Statistics	12	11*	10	9	8
$\bar{R}^2$	0.43	0.44	0.43	0.41	0.34
AIC	6.80	6.79	6.80	6.84	6.95

From Table S3, the appropriate lag length of Y_t is 11. This is due to the highest adjusted R-squared and lowest value of AIC.

S.4. Choosing the Degree of the Polynomial Y_t

Here, we started with a third-degree polynomial and decreased it until we obtained a satisfactory fit.

Table S4: Choosing the degree of the polynomial.

	Equation					
Model Statistics	6th order	5th order	4th order	3rd order	2nd order*	1st order
$\bar{R}^2$	0.42	0.42	0.43	0.43	0.44	0.32
AIC	6.76	6.73	6.71	6.70	6.67	6.86

Therefore, the best order of the polynomial is 2 as shown (starred) in Table S4. This is due to its highest adjusted R-squared and lowest value of AIC. Hence, the best model is ARPDL(11,2,23,6);

$$Y_t = \alpha + \sum_{i=1}^{23} (\gamma_0 i^0 + \gamma_1 i^1 + \gamma_2 i^2 + \cdots + \gamma_6 i^6) X_{t-i} + \sum_{i=1}^{11} (\gamma_0 i^0 + \gamma_1 i^1 + \gamma_2 i^2) Y_{t-i} + \varepsilon_t \quad \text{S.1}$$

Table S5: Results of the autoregressive polynomial distributed lag ARPDL(11,2,23,6) model.

Variable	Coefficient	Std. Error	t-Statistic	p-value
C	-30.11761	9.751959	-3.088366	0.0030
Z_{0t}	0.042895	0.030885	1.388862	0.1699
Z_{1t}	0.002616	0.012444	0.210257	0.8342
Z_{2t}	-0.004550	0.003812	-1.193765	0.2372
Z_{3t}	-0.000839	0.000409	-2.053707	0.0443
Z_{4t}	9.03E-05	8.17E-05	1.105871	0.2731
Z_{5t}	7.20E-06	2.84E-06	2.537704	0.0137
Z_{6t}	-4.45E-07	4.34E-07	-1.025221	0.3093
Z_{7t}	-0.301532	0.116434	-2.589716	0.0120
Z_{8t}	0.032423	0.012380	2.618874	0.0111
Z_{9t}	0.002726	0.003678	0.741171	0.4614

R-squared	0.694957	Mean dependent var	36.09722
Adjusted R-squared	0.644950	S.D. dependent var	8.671078
S.E. of regression	5.166752	Akaike info criterion	6.262129
Sum squared resid	1628.415	Schwarz criterion	6.609953
Log likelihood	-214.4366	Hannan-Quinn criter.	6.400599
F-statistic	13.89721	Durbin-Watson stat	1.922923
Prob(F-statistic)	0.000000		

Note that the created variables from Z_{0t} to Z_{6t} refer to the lag of X_{t-i} whereas the variables from Z_{7t} to Z_{9t} refer to the lag of Y_{t-i}.

The fitted model is shown in Figure S1 together with residual diagnostic plots. This is followed by the distribution of the series in the histogram with a complement of standard descriptive statistics displayed along with the histogram (see Figure S2). The p-value (p=0.60) of the Jarque-Bera test is not less than 0.05 for a 5% significance level and hence we do not reject the null hypothesis that the model is normally distributed. Figure S3 shows leverage plots of the residuals.

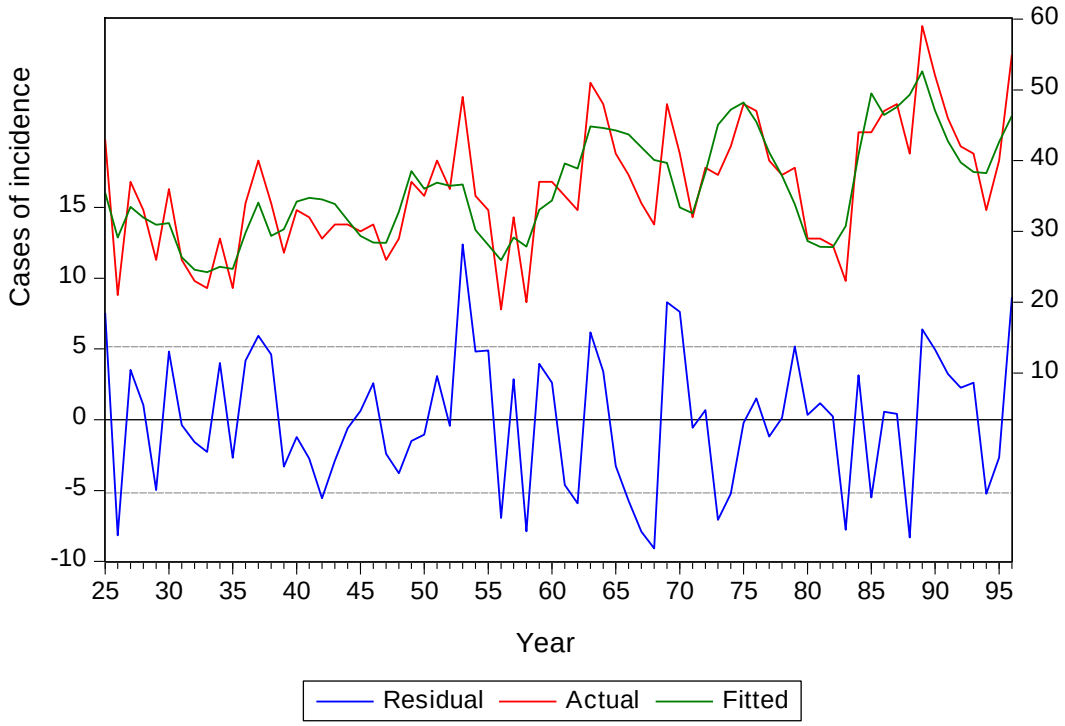


Figure S1: Fitted and residual plots for the best ARPDL(11,2,23,6) model of lung cancer cases per month from 2000 to 2007.

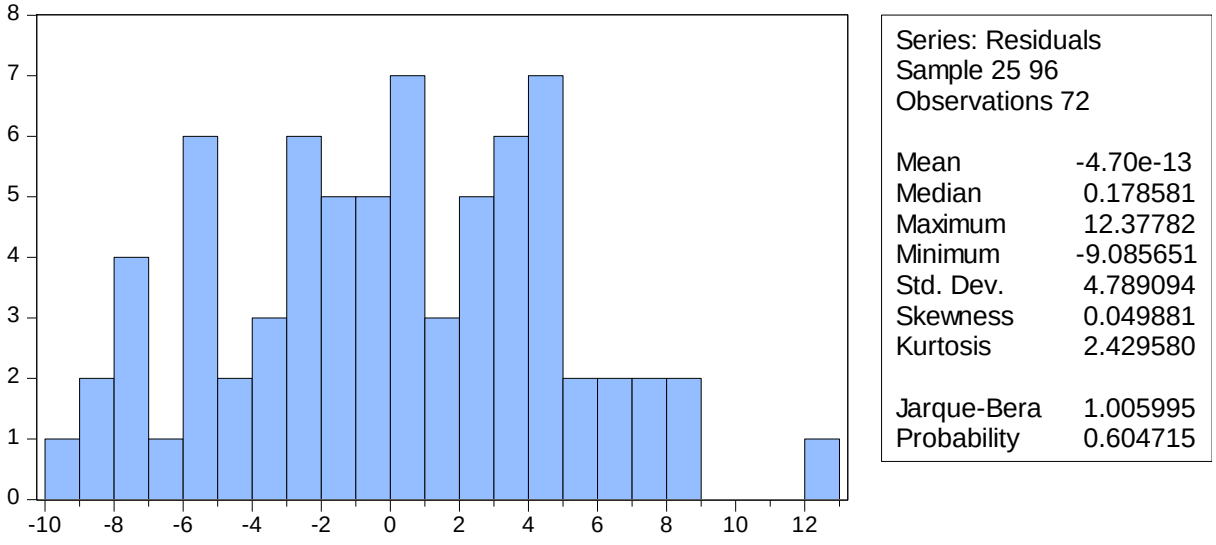


Figure S2: Residual diagnostic of the normality test of the best ARPDL(11,2,23,6) model of lung cancer cases per month from 2000 to 2007.

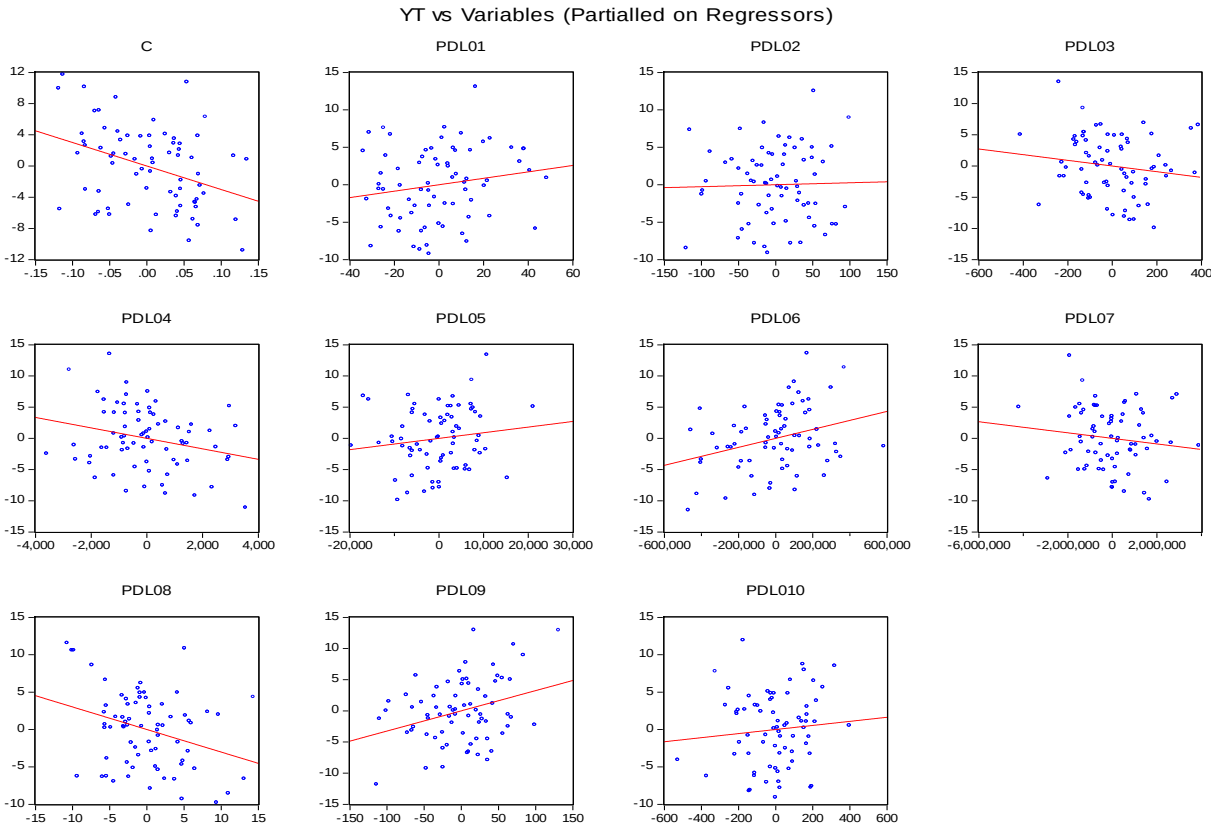


Figure S3: Leverage plots for the stability of diagnostics of the best OLS model of lung cancer cases per month from 2000 to 2007.

S.5. The Breusch-Godfrey Test for Serial Correlation

From Table S6, the values of both the LM-statistic and the F-statistic are low, indicating that we do not reject the null hypothesis and hence conclude that there is no significant serial correlation. Residuals generated from the model are not serially correlated because the p-values are not very small i.e. they are not less than 0.05 for a 5% significance level. Hence, we forecast this model and present the k-step ahead forecast as shown Figure S4.

Table S6: Results of Breusch-Godfrey LM test of ARPDL(11,2,23,6) model.

F-statistic	0.001338	Prob. F(1,84)	0.9709
Obs*R-squared	0.001529	Prob. Chi-Square(1)	0.9688

S.6. Results

The one step ahead out-of-sample forecast was performed on the data from 2008 to 2009 to check the validity of the ARPDL(11,2,23,6) model. Figure S5 shows the actual cases of lung cancer from 2000 to 2009 and the forecast value between 2008 and 2009.

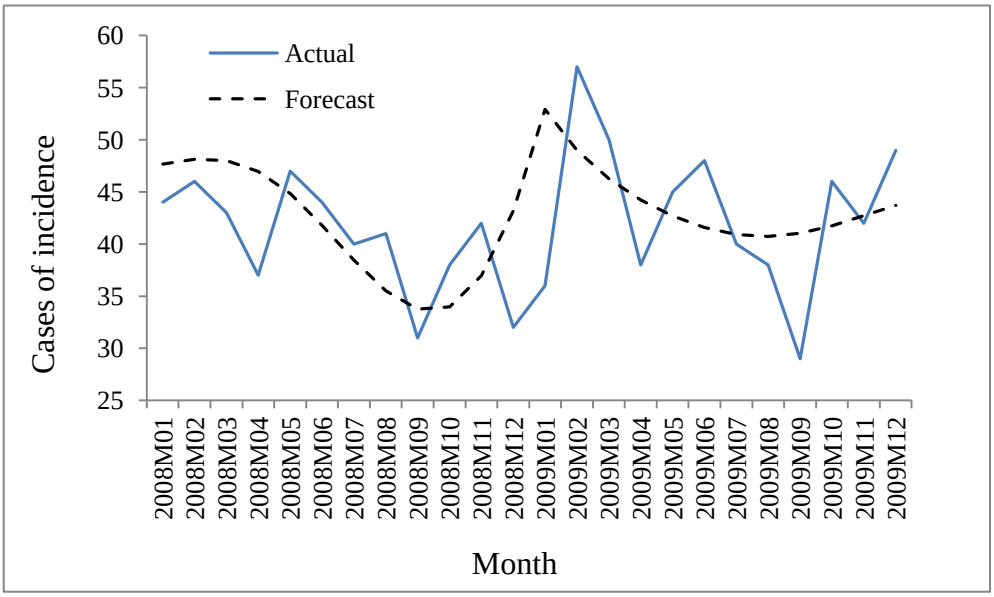


Figure S4: Actual and forecast ARPDL(11,2,23,6) model with 24 months ahead forecast of lung cancer cases per month from 2008 to 2009.

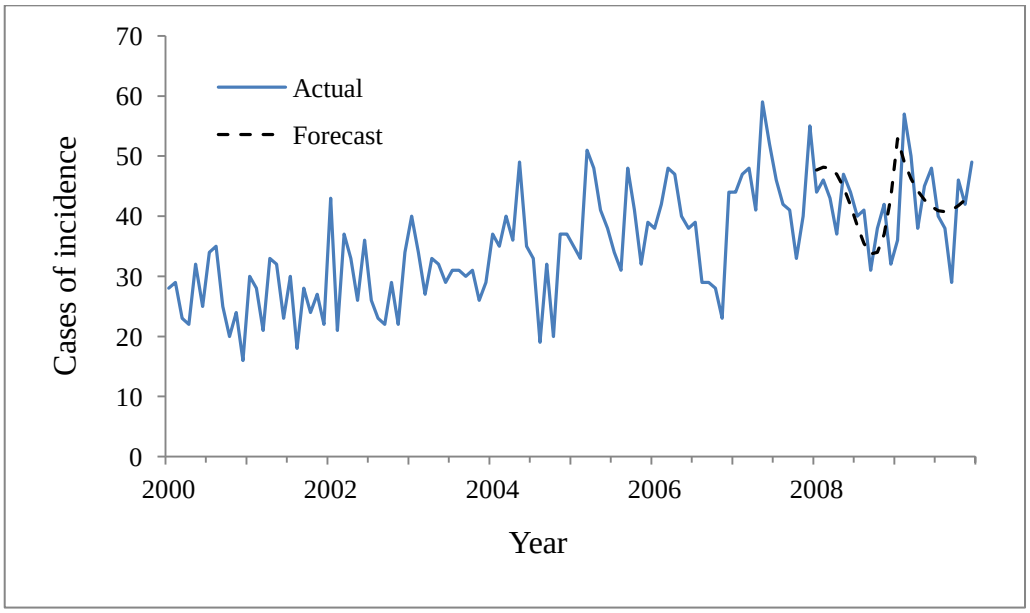


Figure S5: Actual and forecast ARPDL(11,2,23,6) model with 24 months ahead forecast of lung cancer cases per month from 1994 to 2009.

Appendix L: Leverage Plots for the Stability of Diagnostics Check (Model II).

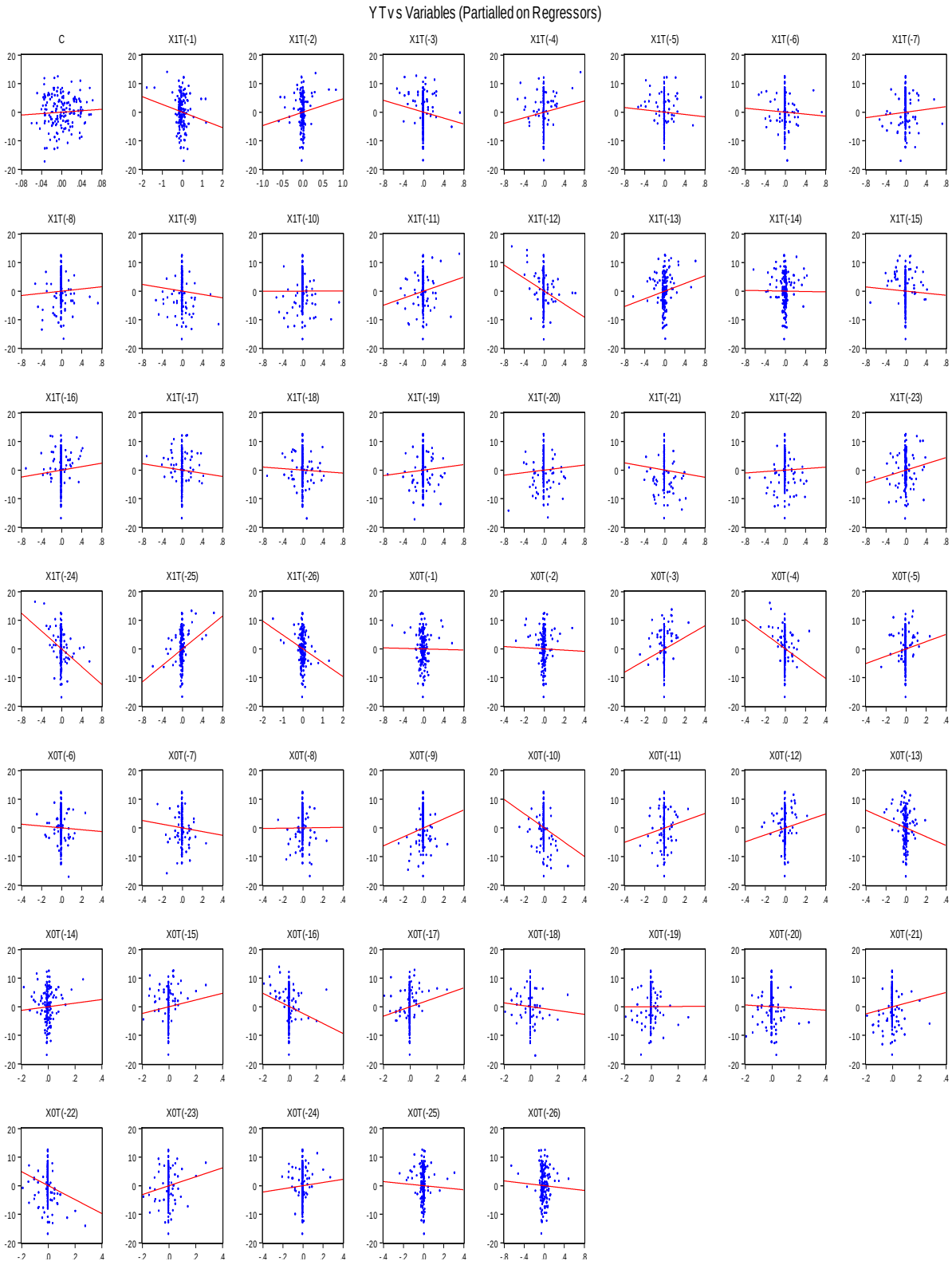


Figure L1: Leverage plots for the stability of diagnostics of the best OLS model of lung cancer cases per month from 1994 to 2009.

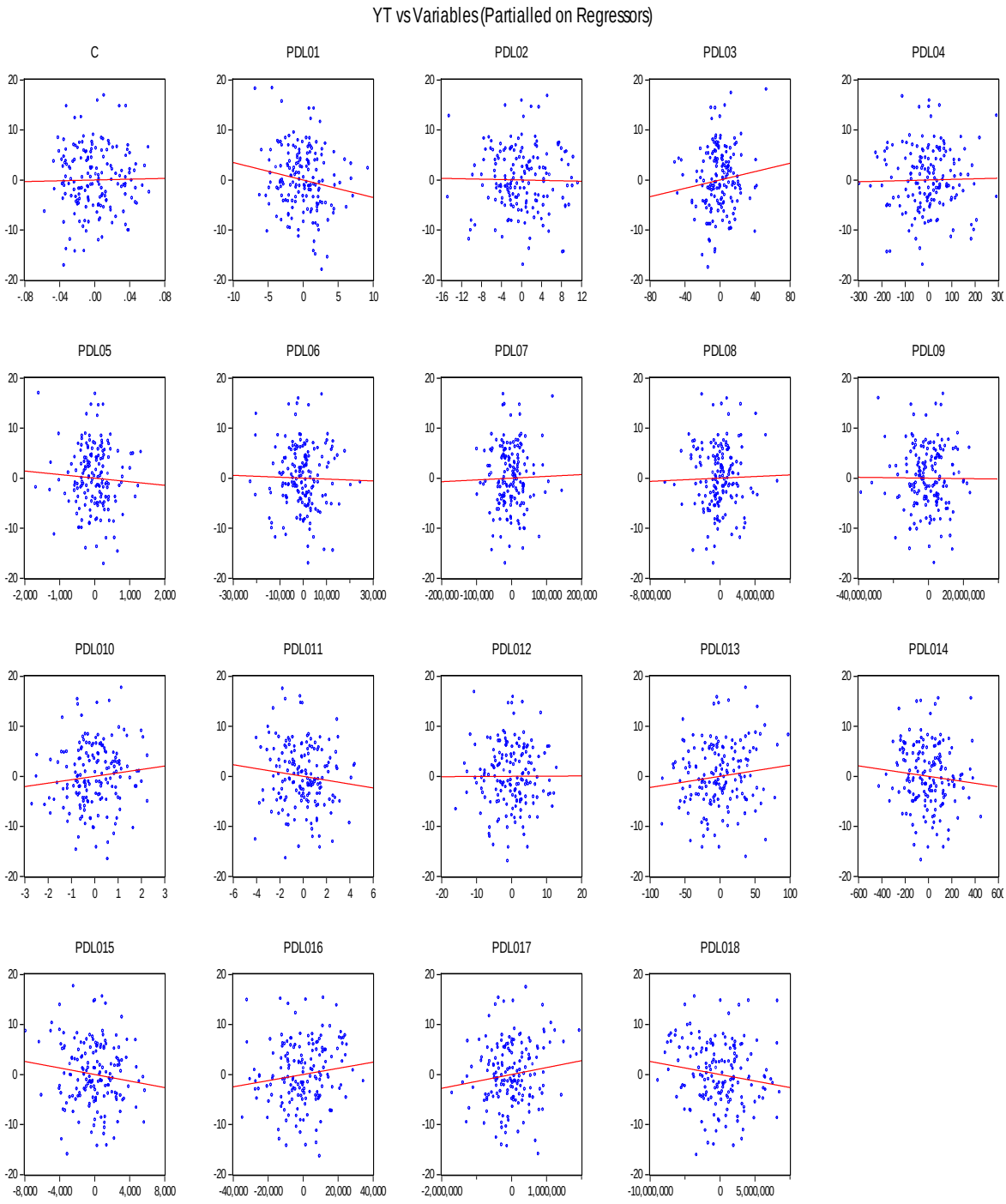


Figure L2: Leverage plots for the stability of diagnostics of the best PDL(26,8) model of lung cancer cases per month from 1994 to 2009.

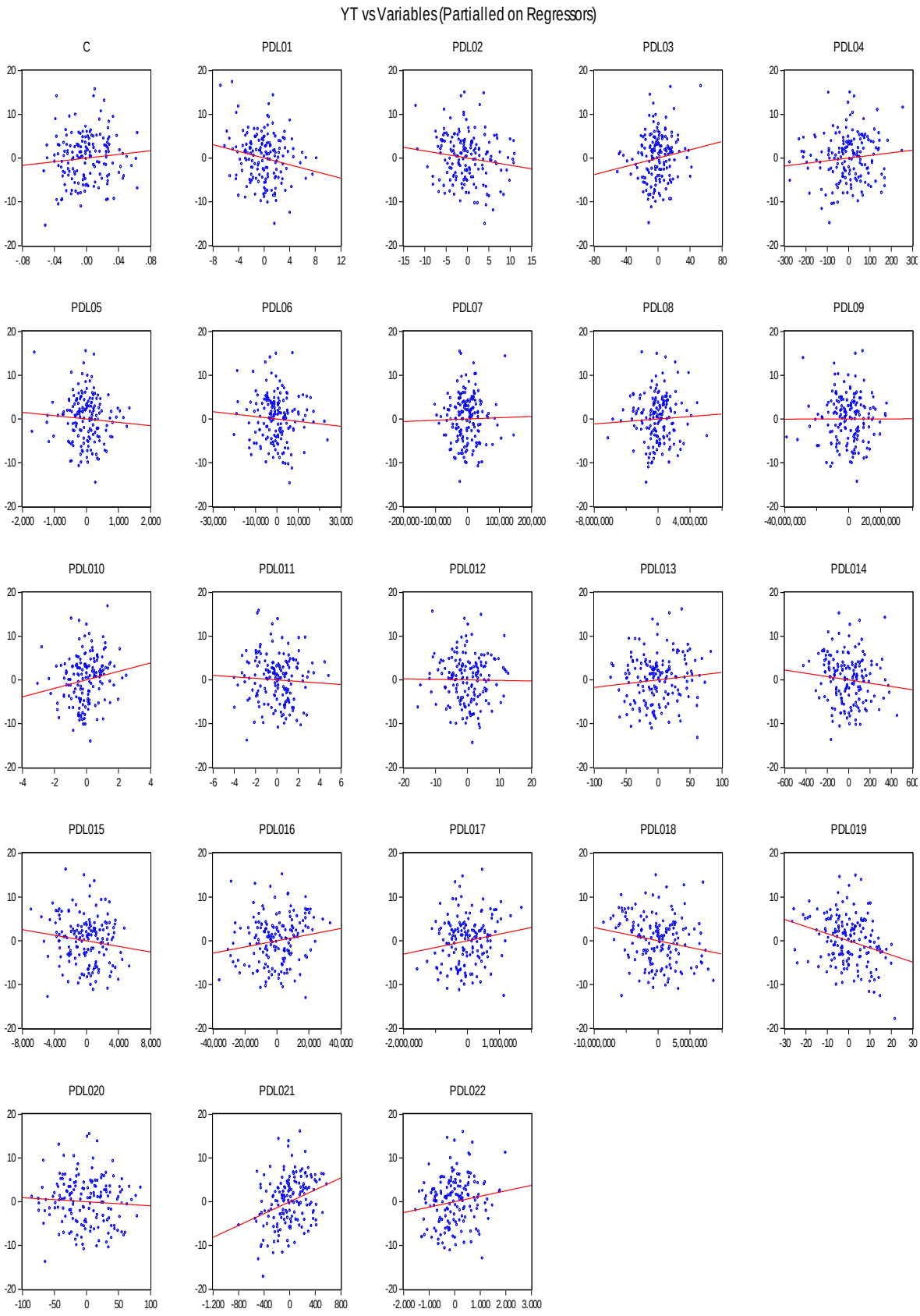


Figure L3: Leverage plots for the stability of diagnostics of the best ARPDL(12,3,26,8) model of lung cancer cases per month from 1994 to 2009.

REFERENCES

Abraham, B., & Ledolter, J. (1983) *Statistical methods for forecasting*. New York: John Wiley & Sons.

Abraham, G., Byrnes, G. B. and Bain, C. A. (2009) Short-Term Forecasting of Emergency Inpatient Flow. *IEEE Transactions on Information Technology in Biomedicine,* 13 (3), 380-388.

Aidoo, E. (2010) Modelling and Forecasting Inflation Rates in Ghana: An Application of SARIMA Models. Master's Thesis. Högskolan Dalarna School of Technology and Business Studies. Sweden.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In: B. N. Petrov and F. Csaki. *Second International Symposium on Information Theory*. Budapest: Akailseoniai-Kiudo, 267-281.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* 19, 716-723.

Al-Ahmadi, K. and Al-Zahrani A. (2011) *Cancer Atlas of Saudi Arabia*. Saudi Arabia: King Abdulaziz City for Science and Technology. Pp.6-8.

Alberg, A.J. and Samet, J.M. (2003) Epidemiology of lung cancer. *Chest 2003*, 123, 21S-49S.

Al-Eid, H.S. (2009) *Cancer Incidence Report. Ministry of Health*. Kingdom of Saudi Arabia: Saudi Cancer Registry.

Almon, S. (1965) The distributed lag between capital appropriations and expenditures. *Econometric*s, 33, 178-196.

American Cancer Society (2011) *Global Cancer Facts & Figures.* 2nd Edition. Atlanta: *American Cancer Society*.

Armstrong, J.S. (1978) Forecasting with Econometric Methods: Folklore versus Fact with Discussion. *Journal of Business*, 51, 549-600.

Asteriou, D. and Hall, G.S. (2011) *Applied econometrics*. China: Palgrave Macmillan.

Baker, A. and Bray, I. (2005) Bayesian projections: What are the effects of excluding data from younger age groups? *Am. J. Epidemiol,* 162, 798–805.

Barretto, H. and Howland, F.M. (2006) *Introductory Econometrics*. U.S.A: Cambridge University Press, New York.

Bentzen, J. and Engsted, T. (2001) A revival of the autoregressive distributed lag model in estimating energy demand relationships. *Energy,* 26, 45-55.

Berzuini, C. and Clayton, D. (1994) Bayesian analysis of survival on multiple time scales. *Stat. Med,* 13, 823–838.

Berzuini, C., Clayton, D. & Bernardinelli, L (1993) Bayesian inference on the Lexis diagram. *Bulletin of the International Statistical Institute*, 55 (1), 149-165.

Besag, J., Green. J. P., Higdon, D. and Mengersen, K. (1995) Bayesian computation and stochastic systems (with discussion). *Statistical Science,* 10 (1), 3–66.

Bhansali, R.J. (1996) Asymptotically efficient autoregressive model selection for multistep prediction. *Annals of the Institute of Statistical Mathematics*, 48, 577-602.

Bhansali, R.J. (1999) Autoregressive model selection for multistep prediction. *Journal of Statistical Planning and Inference,* 78, 295–305.

Bowerman, Bruce. L., Richard, T. O'Connell, and Anne, B. Koehler. (2005) *Forecasting, Time Series, and Regression*. 4th Edition. Belmont, CA: Thomson Brooks/Cole.

Box, G. and G. Jenkins (1976) *Time Series Analysis: Forecasting and Control. Revised Edition*. Holden Day.

Box, G. E., Jenkins, G. M. & Reinsel, G. C. (2008) *Time Series Analysis: Forecasting and Control.* New York: John Wiley and Sons.

Box, G.E.P., & Jenkins, G.M. (1970) *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day (revised ed. 1976).

Box, George., Gwilym, M., Jenkins, and Gregory, C. (1994) *Time Series Analysis: Forecasting and Control*. Third edition. Prentice Hall.

Boyle, P. and Levin, B. (2008) *World Cancer Report*. Lyon: International Agency for Research on Cancer.

Bray, F. and Moller, B. (2006) Predicting the future burden of cancer. *Nat Rev Cancer,* 6, 63-74.

Bray, I. (2002) Application of Markov chain Monte Carlo methods to projecting cancer incidence and mortality. *J. Roy. Statist. Soc. Ser. C,* 51, 151–164.

Bray, I., Brennan, P. and Boffetta, P. (2001) Recent trends and future projections of lymphoid neoplasm: a Bayesian age-period-cohort analysis. *Cancer Causes Control*, 12, 813–820.

Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models. *Journal of American Statistical Association*, 88, 9-25.

Breusch, T.S. (1978) Testing for autocorrelation in dynamic linear models. *Aurtral. Econom. Papers*, 17, 334-355.

Brown, R.G. (1959) *Statistical forecasting for inventory control*. New York: McGraw-Hill.

Brown, R.G. (1963) *Smoothing, forecasting and prediction of discrete time series*. Englewood Cliffs, NJ: Prentice-Hall.

Burnham, K.P., & Anderson, D.R. (1998) *Model Selection and Inference*. New York: Springer-Verlag.

Byers, A.L., Allore, H., Gill, T.M. & Peduzzi, P.N. (2003) Application of negative binomial modelling for discrete outcomes: A case study in aging research. *Journal of Clinical Epidemiology*, 56, 559-564.

Cagan, P. (1956) *The monetary dynamics of hyperinflation*. In M. Friedman (Ed.), Studies in the Quantity Theory of Money. Chicago: University of Chicago Press.

Carstensen, B. (2007) Age–period–cohort models for the Lexis diagram. *Statistics in Medicine*, 26, 3018–3045.

Chatfield, C. (2004) *The Analysis of Time Series: An Introduction*. Chapman and Hall/CRC.

Clayton, D. and Schifflers, E. (1987a) Models for temporal variation in cancer rates I: age-period and age-cohort models. *Statistics in Medicine,* 6, 449-467.

Clayton, D. and Schifflers, E. (1987b) Models for temporal variation in cancer rates II: age-period-cohort models. *Statistics in Medicine,* 6, 469-481.

Clements, M. S., Armstrong, B. K. and Moolgavkar, S. H. (2005) Lung cancer rate predictions using generalized additive models. *Biostatistics*, 6, 576-589.

Cleries, R., Martinez, J. Escriba, J., Esteban, L., Pareja, L., Borras, J., & Ribes, J. (2010) Monitoring the decreasing trend of testicular cancer mortality in Spain during 2005-2019 through a Bayesian approach. *Cancer Epidemiology*, 135, 0-13.

Cleries, R., Martinez, J. M., Valls, J., Pareja, L., Esteban, L., Gispert, Moreno, Ribes, R. V. J. and Borras, J. M. (2009) Life expectancy and age-period- cohort effects: analysis and projections of mortality in Spain between 1977 and 2016. *Public Health*, 123 (2), 156-162.

Cleries, R., Ribes, J., Esteban, L., Martinez, J.M., and Borras, J.M. (2006) Time trends of breast cancer mortality in Spain during the period 1977−2001 and Bayesian approach for projections during 2002−2016. *Ann Oncol*, 17(12), 1783−1791.

Cochrane, D. and Orcutt, G. (1949) Application of least squares regression to relationships containing autocorrelated error terms. *Journal of the American Statistical Association*, 44, 32-61.

Cochrane, J. H. (2005b) Time series for macroeconomics and finance. *Manuscript, University of Chicago.*

Cochrane, John. H. (1997) Time Series for Macroeconomics and Finance. *Unpublished book manuscript.*

Congdon, P. (2006) *Bayesian Statistical Modelling*. 2nd Edition. England: John Wiley & Sons Ltd.

Cooper, P. J. (1972) Two approaches to polynomial distributed lags estimation: an expository note and comment. *American Statistician,* 26, 32–35.

Cooper, R.L. (1972) *The Predictive Performance of Quarterly Econometric Models of the United States.* New York: National Bureau of Economic Research.

Davidson, R. and MacKinnon, J. (1993) *Estimation and Inference in Econometrics*. New York: Oxford University Press.

De Vries V.M. (1927) The prevalence of cancer as revealed by mortality returns and at autopsy. *Surg Gyn Obst*, 217-245.

del Moral, M.J. & Valderrama, M. J. (1997) A principal component approach to dynamic regression models. *International Journal of Forecasting*, 13, 237–244.

Devesa, S. S., Silverman, D. T., Young, J. L., Pollack, E. S., Brown, C. C., Horm, J. E. (1987) Cancer incidence and mortality trends among whites in the United States, 1947-84. *J Natl Cancer Inst*, 79, 701-770.

Doll, R., Payne, P. and Waterhouse, J. (1966) *Cancer Incidence in Five Continents*. Geneva, UICC. Berlin: Springer, Volume 1.

Dos Santos Silva I. (1999) *Cancer Epidemiology: Principles and Methods*. Lyons: International Agency for Research on Cancer.

Durbin, J. (1970) Testing for serial correlation in least-squares regressions when some of the repressors are lagged dependent variables. *Econometrica*, 38, 410–421.

Durbin, J. and Watson, G.S. (1950) Testing for serial correlation in least-squares regression, I, *Biometrika*, 37, 409-428.

Durbin, J., & Koopman, S. J. (2001) *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

Edlund, P. O. (1984) Identification of the multi-input Box-Jenkins transfer function model. *Journal of Forecasting*, 3, 297–308.

Elkum, N. (2005) Prediction confidence intervals for the age-period cohort model. *Journal of Data Science*, 3, 403-414.

Elliot, J.W. (1973) A Direct Comparison of Short-Run GNP Forecasting Models. *Journal of Business*, 46, 33-60.

Elsayed, I. S, Abdul, R.J, and Malcolm, A.M. (2011) Lung cancer incidence in the Arab League countries. *Asian Pacific Journal of Cancer Prevention*, 12, 17-34.

Faraway, J. & Chatfield, C. (1998) Time series forecasting with neural networks: a comparative study using the airline data. *Applied Statistics*, 47, 231–250.

Ferlay J, Shin, H.R, Bray F, Forman D, Mathers C.D. and Parkin D. (2010) *GLOBOCAN 2008, Cancer Incidence and Mortality Worldwide:* IARC CancerBase No.10 [Internet]. Lyon, France: International Agency for Research on Cancer. Available from: http://globocan.iarc.fr.

Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D. and Bray, F. (2013) *GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide*: IARC CancerBase No. 11 [Internet].Lyon, France: International Agency for Research on Cancer. Available from: http://globocan.iarc.fr, accessed on 15/03/2015.

Fienberg, S. E. and Mason, W. M. (1985) *Specification and implementation of age, period, and cohort models*. Pp.45–88 in cohort analysis in social research, edited by William M. Mason and Stephan E. Fienberg. New York: Springer-Verlag.

Fienberg, S.E. and Mason, W.M. (1978) Identification and estimation of period-age-cohort effects in the analysis of discrete archival data, *Sociological methodology1979*, 1-67.

Fomby, T.B. Hill, R.C. and Johnson, S. R. (1984) *Advanced Econometric Methods*. New York: Springer.

Franses, P.H and Oest, R. D. (2004) *On the econometrics of the Koyck model* (No. EI 2004-07). Economic Institute Research Papers. Erasmus University Rotterdam.

Fu, W. J. (2000) Ridge estimator in singular design with application to age-period-cohort analysis of disease rates. *Communications in Statistics-Theory and Method*, 29, 263-78.

Fu, W. J., Peter, H. and Thomas, E. R. (2004) Age-period-cohort analysis: structure of estimators, estimability, Sensitivity, and asymptotics. *Journal of the American Statistical Association*, revised and resubmitted.

Galbraith, J. W. & Zinde-Walsh, V. (2001) *Autoregression-based estimators for ARFIMA models* (No. 2011s-11.). CIRANO.

Gelman, A. (2005) Prior distributions for variance parameters in hierarchical models. *Bayesian Anal*, 1, 1–19.

Gelman, A. and Rubin, D. (1992) A single series from the Gibbs sampler provides a false sense of security. *Bayesian Statistics, 4*, 625–631.

Gelman, A. and Rubin, D. (1992a) Inference from Iterative Simulation using Multiple Sequences. *Statistical Science*, 7, 457–511.

Glenn, N. D. (1976) Cohort analysts' futile quest: statistical attempts to separate age, period, and cohort effects. *American Sociological Review*, 41, 900–904.

GLIM. (1986). *Generalised Linear Interactive Modelling. Release* 3.77 Manual. New York.

Godfrey, L.C. (1978) Testing against general autoregressive and moving average error models when the repressors include lagged dependent variables, *Econometrics*, 46, 1293-1302.

Goldberger, A.S. (1991) *A Course in Econometrics*. Cambridge, MA: Harvard University.

Greene, W.H. (2000) *Econometric Analysis*. New York: Prentice Hall.

Gulf Cooperation Council (GCC) (2011) *Ten year cancer incidence among nationals of the GCC states, 1998-2000.* Kingdom of Saudi Arabia: National Cancer Registry. Pp.2-3.

Hamzacebi, C. (2008) Improving artificial neural networks' performance in seasonal time series forecasting. *Information Sciences*, 178, 4550-4559.

Harrison, P.J., & Stevens, C.F. (1976) Bayesian forecasting. *Journal of the Royal Statistical Society (B)*, 38, 205–247.

Harvey, A.C. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Harvey, A.C. (2006) Forecasting with unobserved component time series models. In Elliot, G., C.W.J. Granger & A. Timmermann (eds.). *Handbook of Economic Forecasting*, Amsterdam: Elsevier Science.

Heaton, M. J. and Peng, R.D. (2013) Extending distributed lag models to higher degrees. *Biostatistics*, **0** (0), 1-23.

Hendry, D., Pagan, A. and Sargan, J. (1984) *Dynamic Specifications*. 2nd ed. North Holland: Amsterdam.

Hilderth, C. and Lu, J. (1960) *Demand Relations with Autocorrelated Disturbances*. Technical Bulletin No. 276. Michigan State University Agricultural Experiment Station.

Hill, C., Griffiths, W. and Judge, G. (2000) *Undergraduate Econometrics*. 2nd ed. New York: John Wiley and Sons.

Hipel, K. W. and McLeod, A. I. (1994) *Time Series Modelling of Water Resources and Environmental Systems*. Amsterdam: Elsevier.

Hipel, K.W. and McLeod, A.I. (2005) *Time Series Modeling of Water Resources and Environmental Systems.* Electronic reprint of our book originally published in 1994.

Hobcraft, J., Menken J. and Preston S. (1982) Age, period, and cohort effect in demography. *Population Index*, 48, 4–43.

Holford, T. R. (1983) The estimation of age, period and cohort effects for vital rates. *Biometrics*, 39, 311–324.

Holford, T. R. (1985) An alternative approach to statistical age-period-cohort analysis. *Journal of Clinical Epidemiology*, 38, 831-836.

Holford, T. R. (1991) Understanding the effects of age, period, and cohort on incidence and mortality rates. *Annual Review of Public Health*, 12, 425–57.

Holford, T. R. (1992) Analysing the temporal effects of age, period, and cohort. *Statistical methods in medical research*, 1, 317–37.

Holt, C.C. (1957) Forecasting seasonals and trends by exponentially weighted averages. O.N.R. Memorandum 52/1957. Carnegie Institute of Technology. Reprinted with discussion in 2004. *International Journal of Forecasting*, 20, 5–13.

Huang, Y., Dominici, F. and Bell, M. (2004) Bayesian hierarchical distributed lag models for summer ozone exposure and cardio-respiratory mortality. *Technical report, John Hopkins University, Dept. of Biostatistics*.

Hurvich, C.M., & Tsai, C.L (1989) Regression and Time Series Model Selection in Small Sample. *Biometrika*, 76, 297-307.

Hyndman, R. and Khandakar, Y. (2008) Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software*, 27(3), 1-22.

IARC. (2013) Latest world cancer statistics. [Press release]. [Accessed 10 Feb 2015]. Available from http://www.iarc.fr/en/media-centre/pr/2013/pdfs/pr223_E.pdf

International Agency for Research on Cancer. (2014). World cancer factsheet. [Online]. [Accessed 10 Jan 2015]. Available from http://publications.cancerresearchuk.org/downloads/product/CS_REPORT_WORLD.pdf.

Jeffry, M. W. (2003) *Introduction to econometrics*. United State of America: South-Western, a division of Thomson Learning.

Kalman, R.E. (1960) A new approach to linear filtering and prediction problems. Transaction of the ASME. *Journal of Basic Engineering*, 82D, 35–45.

Kaplan, D. (2014) *Bayesian Statistics for the Social Sciences*. The Guilford Press. NY, USA.

Khellaf, M., Quantin, C., d'Athis, P., Fassa, M., Jooste, V., Hervieu, M., Giroud, M., Be´jot, Y. (2010) Age–period–cohort analysis of stroke incidence in Dijon from 1985 to 2005. *Journal American Heart Association*, 41, 2762-2767.

Kleiber, C. and Zeileis, A. (2008) *Applied Econometrics with R*. New York: .Springer-Versa.

Kmenta, J. (1986) *Elements of Econometrics*. New York: Macmillan.

Knight, K. and Fu, W. J. (2000) Asymptotics for Lasso-Type estimations. *Annals of Statistics*, 28, 1356–78.

Knorr-Held, L. and Rainer, E. (2001) Projections of lung cancer mortality in West Germany: A case study in Bayesian prediction. *Biostatistics*, 2, 109–129.

Kolmogorov, A.N. (1941) Stationary sequences in Hilbert space (Russia). *Bull. Moscow State Univ. Math*, p.40.

Koyck, L.M. (1954) *Distributed lags and investment analysis*. Amsterdam: North-Holland.

Krishnamurthi, L., Narayan, J. & Raj, S.P. (1989) Intervention analysis using control series and exogenous variable in a transfer function model: A case study. *International Journal of Forecasting*, 5, 21–27.

La, V. C., Lucchini, F., Negri, E., Boyle, P., Maisonneuve, P., Levi, F. (1992) Trends of cancer mortality in Europe, 1955-1989: II, respiratory tract, bone, connective and soft tissue sarcomas, and skin. *Eur J Cancer*, 28, 514-599.

Ledolter, J., & Abraham, B. (1984). Some comments on the initialization of exponential smoothing. *Journal of Forecasting*, 3, 79-84.

Lee, R. D. and Carter, L. R. (1992) Modeling and forecasting U.S. mortality. *J. Amer. Statist. Assoc*, 87, 659–671.

Lee, T. C. K., Dean, C. B. and Semenciw, R. (2011) Short-term cancer mortality projections: a comparative study of prediction methods. *Statistics in Medicine*, 30, 3387-3402.

Maddala, G.S. (1977) *Econometrics*. McGraw-Hill: Singapore.

Maddala, G.S. and Lahiri, K. (2009) *Introduction to Econometrics*. 4th ed. Glasgow: UK.

Maeshiro, A. (1996) Teaching regressions with a lagged dependent variable and autocorrelated disturbances. *The Journal of Economic Education*, 27(1), 72–84.

Makridakis, S., Wheelwright, S.C. & Hyndman, R.J. (1998) *Forecasting Methods and Applications*. 3rd Edition. New York: Wiley.

Mason, K. O., Mason, W. M., Winsborough, H. H., & Poole, W. K. (1973) Some methodological issues in cohort analysis of archival data. *American Sociological Review*, 38, 242-258.

Mason, K. O., William, H. Mason, H. H., Winsborough, and Poole, W. K. (1973) Some methodological issues in cohort analysis of archival data. *American Sociological Review*, 38, 242–58.

Mason, W. M. and Smith, H. L. (1985) *Age-period-cohort analysis and the study of deaths from pulmonary tuberculosis*. New York: Springer-Verlag.

McCullagh, P. & Nelder, J.A. (1983) *Generalized linear models*. New York: Chapman and Hall.

McWhorter, A. Jr. (1975) Time Series Forecasting Using the Kalman Filter: An Empirical Study. *Proceedings of the American Statistical Association: Business and Economics Section*, 436-446.

Midorikawa, S., Miyaoka, E. and Smith, B. (2008) Application of dynamic Poisson models to Japanese cancer mortality data. *Journal of Modern Applied Statistical Methods*, 7(2), 22-23.

Mistry, M., Parkin, D. M., Ahmad, A. S. and Sasieni, P. (2011) Cancer incidence in the United Kingdom: projections to the year 2030. *Br J Cancer*, 105, 1795-1803.

Moller, B., Fekjr, H., Hakulinen, T., Sigvaldason, H., Storm, H. H. M., Talback, and Handorsen, T. (2003) Prediction of cancer incidence in the Nordic countries: empirical comparison of different approaches, *Statistics in Medicine*, 22, 2751-2766.

Montgomery, D.C., Peck, E.A. & Vining, G.G. (2006) *Introduction to linear regression analysis*. New York: John Wiley and Sons.

Muth, J.F. (1960) Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association*, 55, 299–306.

Nakamura, T. (1986) Bayesian cohort models for general cohort table analyses. *Ann. Inst. Statist. Math.* 38, 353–370.

Narasimham G. V., Castellion, V. F. and Singpurwalla, N.D. (1974) On the Predictive Performance of the BEA Quarterly Econometric Model and a Box-Jenkins Type ARIMA Model. *Proceedings of the American Statistical Association*, 2, 501-504.

Nelson, C.R. (1972) The Prediction Performance of the FRB-MIT-PENN Model of the US Economy. *American Economic Review,* 5, 902-917.
Nerlove, M. (1958) *The Dynamic of supply: Estimation of Farmer's Response to price*. Johns Baltimore: Hopkins University Press.
Nerlove, M. (1958b) Distributed lags and estimation of long-run supply and demand elasticties: theoretical considerations, *Journal of Farm Economics 40(2),* 301-314.
Nerlove, M. (1959) Distributed lags and demand analysis for agricultural and other commodities. *Journal of Farm Economics*, 41(1), 151-153.
Newbold, P. (1983) ARIMA model building and the time series analysis approach to forecasting. *Journal of Forecasting*, 2, 23–35.
O'Brien, R. M. (2000) Age-period-cohort characteristic models. *Social Science Research*, 29, 123–139.
O"Donovan, T, M. (1983) *Short Term Forecasting: An introduction to the Box-Jenkins Approach*. New York: John Wiley & Sons.
Osmond, C. (1985) Using age, period and cohort models to estimate future mortality rates. *International Journal of Epidemiology*, 14, 124–129.
Osmond, C., and Gardner, M. J. (1982) Age, period, and cohort models applied to cancer mortality rates. *Statistics in Medicine*, 1, 245–259.
Pankratz, A. (1991) *Forecasting with dynamic regression models.* New York: John Wiley and Sons.
Park, H. (1999) Forecasting Three-Month Treasury Bills Using ARIMA and GARCH Models. *Econ*.
Pegels, C.C. (1969) Exponential smoothing: some new variations. *Management Science*, 12, 311–315.
Pena, D., & Sanchez, I. (2005) Multifold predictive validation in ARMAX time series models. *Journal of the American Statistical Association*, 100, 135–146.
Prais, S. J. and Winsten, C. B. (1954) *Trend estimators and serial correlation*. Vol. 383, pp. 1-26. Chicago: Cowles Commission discussion paper.
Quenouille, M. H. (1957) *The Analysis of Multiple Time-Series*, London: Griffin. (2nd ed. 1968).
Quinn, M. J., Babb, P. J., Brock, A., Kirby, E. A. and Jones, J. (2001) *Cancer trends in England and Wales, 1950-1999.* Studies on medical and population subjects No.66. London: The Stationery Office.
Quinn, M. J., d'Onofrio, A., Møller, B., Black, R., Martinez-Garcia, C., Møller, H., Rahu, M., Robertson, C., Schouten, L. J., La Vecchia, C. and Boyle1, P. (2003) Cancer mortality trends in the EU and acceding countries up to 2015. *Ann Oncol*, 14, 1148-1152.
Raftery, A. E. (1995) Bayesian model selection in social research. *Sociological methodology*, 25, 111-164.
Raicharoen, T., Lursinsap, C. and Sanguanbhokai, P. (2003) Application of critical support vector machine to time series prediction. In *Circuits and Systems, 2003. ISCAS'03. Proceedings of the 2003 International Symposium on* (Vol. 5, pp. V-741). IEEE.
Raifu, A. and Arbyn, M. (2009) Description of cervical cancer mortality in Belgium using Bayesian age-period-cohort models. *Arch Public Health*, 67, 100-115.
Roberts, S.A. (1982) A general class of Holt-Winters type forecasting models. *Management Science*, 28, 808–820.
Robertson, C. and Boyle, P. (1998) Age-period-cohort analysis of chronic disease rates. I: Modelling Approach. *Statistics in Medicine*, 17, 1305-1323.
Robertson, C., Gandini, S., and Boyle, P. (1999) Age-period-cohort models: a comparative study of available methodologies. *Journal of Clinical Epidemiology*, 52, 569–583.

Rue, H., Martino, S. and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Method*, 71, 319–392.

Rutherford, M. J., Lambert, P. C., and Thompson, J. R. (2010) Age–period–cohort modelling. *Stata Journal*, 10, 606–627.

Rutherford, M. J., Thompson, J. T., Lambert P. C. (2012) Projecting cancer incidence using age-period-cohort models incorporating restricted cubic splines. *The International Journal of Biostatistics*, 8:33.

Sadowski, E.A. (2010) A Time Series Analysis: Exploring the Link between Human Activity and Blood Glucose Fluctuation.

Sakamoto, Y., Ishiguro, M., and Kitagawa G. (1986) Akaike Information Criterion Statistics. *Dordrecht, The Netherlands: D. Reidel*.

Sasco, A. J. (1991) World burden of tobacco-related cancer. *Lancet*, 338, 123-4.

Sasieni, P. D. (2012) Age–period–cohort models in Stata. *Stata Journal*, 12, 45-60.

Sasieni, P. D. and Adams J. (1999) Effect of screening on cervical cancer mortality in England and Wales: analysis of trends with an age period cohort model. *British Medical Journal*, 318, 1244–1245.

Sasieni, P. D. and Adams, J. (2000) Analysis of cervical cancer mortality and incidence data from England and Wales: evidence of a beneficial effect of screening. *Journal of the Royal Statistical Society Series A*, 163, 191–209.

Schmid, V. J. and Held, L. (2007) Bayesian age-period-cohort modelling and prediction BAMP. *Journal of Statistical Software*, 21, 1–15.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

Schwarz, J.C., Spix, G., Touloumi, L., Bacharova, T., Barumamdzadeh, A., Le Tertre, T., Piekarski, A., Ponce De Leon, A., Ponka, G., Rossi, M., Saez, and Schouten, J. (1996) Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. *Journal of Epidemiology and Community Health*, 50, 3-11.

Scottish Intercollegiate Guidelines Network. (2005) *Management of patients with lung cancer*. Edinburgh: Clinical Guideline.

Searle, S. R. (1971) *Linear Models*. New York: Wiley.

Shiller, R. J. (1973) A distributed lag estimator derived from smoothness priors. *Econometrics*, 41, 775–788.

Shumway, R. H. and Stoffer, D. (2000) *Time Series Analysis and Its Applications*. New York: Springer-Verlag.

Slutsky, E. (1937) The Sommation of Random Causes as the Source of Cyclic Processes. *Econometrica*, 5, 105-146.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and Van, der Linde AJ. (2002) Bayesian measures of model complexity and fit (with discussion). *J R Stat Soc Ser B*, 64, 583–639.

Stegmueller, D. (2014) Bayesian hierarchical age-period-cohort models with time structured effects: an application to religious voting in the US, 1972-2008. *Electoral Studies*, 33, 52–62.

Stewart, B. W. and Kleihues, P. (2003) *World Cancer Report*. Lyon: IARC Press.

Sverdrup, E. (1967) *Statistics method*. Statistical Memoirs. Institute of Mathematics, University of Oslo (in Norwegian).

Tarone, R. and Kenneth C. Chu. (2000) Age-period-cohort analysis of breast-, ovarian-endometrialand cervical-cancer mortality rates for caucasian women in the USA. *Journal of Epidemiology and Biostatistics*, 5, 221–231.

Tarone, R. and Kenneth, C. Chu (1992) Implications of birth cohort patterns in interpreting trends in breast cancer rates. *Journal of National Cancer Institute*, 1402–1410.

Team, R. C.(2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2015. URL https://www.R-project.org/.

Thomas, R. L. (1997) *Modern Econometrics*. England. Longman: Group United Kingdom.

Tong, H. (1983) *Threshold Models in Non-Linear Time Series Analysis*. New York: Springer-Verlag.

United Nations, Department of Economic and Social Affairs, Population Division. World Population Prospect (2012), accessed date 01-01-2015, available at < http://populationpyramid.net/saudi-arabia/2020/>

Wei, W. W. S. (1990) *Time Series Analysis: Univariate and Multivariate Methods.* California: Addison-Wesley Publishing Company.

Welty, L. Zeger, S. (2005) A sensitivity analysis using flexible distributed lag models. *American Journal of Epidemiology*, 162, 80-88.

West, M., & Harrison, P.J. (1989) *Bayesian Forecasting and Dynamic Models*. New York: Springer-Verlag. (2nd ed., 1997).

Wilmoth, J. R. (1990) Variation in vital rates by age, period, and cohort. *Sociological Methodology*, 295-335.

Winters, P.R. (1960) Forecasting sales by exponentially weighted moving averages. *Management Science*, 6, 324–342.

Wold, H. (1938) *A Study in the Analysis of Stationary Time Series*. Stockholm: Almgrist & Wiksell.

World Health Organization, Media Center, Cancer, (2014), accessed date 4-11-2014, available at <http://www.who.int/mediacentre/factsheets/fs297/en/>

World Health Organization, Media Center, Cancer, (2015), accessed date 24-02-2015, available at < http://www.who.int/mediacentre/factsheets/fs297/en/>

World Health Organization. (2007) *Ten statistical highlights in global public health. World Health Statistics 2007*. Geneva: World Health Organization.

Yaffee, R. & McGee, M. (2000) *Introduction to Time Series Analysis and Forecasting with Applications of SAS and SPSS.* Orlando, Florida: Academic Press.

Yang, Y. and Land, K. C. (2008) Age-period-cohort analysis of repeated cross-section surveys - fixed or random effects? *Sociological Methods and Research*, 36(3), 297–326.

Yang, Y., Fu. J. W., and Land, C. K. (2004) A methodological comparison of age-period-cohort models: The intrinsic estimator and conventional generalized linear models. *Sociological Methodology*, 34, 75−110.

Yule, G.U. (1926) Why Do We Sometimes Get Nonsense-Correlations between Time Series? A Study in Sampling and the Nature of Time Series. *Journal of Royal Statistical Society*, 89, 1-64.

Yule, G.U. (1927) On the method of investigating periodicities in disturbed series, with special reference to Wolfers sunspot numbers. *Philosophical Transactions of the Royal Society London, Series A*, 226, 267–298.

Zhang, G. P. (2007) A neural network ensemble method with jittered training data for time series forecasting. *Information Sciences*, 177, 5329–5346.

Zhang, G.P. (2003) Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50, 159–175.